



TUNIS BUSINESS SCHOOL  
UNIVERSITY OF TUNIS



End of Studies Project

In regard of obtaining the degree of

# Bachelor of Science in Business Administration

Business Analytics

Entitled:

---

**Implementation of a BI solution to  
monitor the commercial pipeline**

---

Presented by:

Ahmed Dammak

Supervised by:

Mr. Aymen Ayari  
Mrs. Fedya Talmoudi

Professional Supervisor  
Academic Supervisor

School Year 2023-2024

# Acknowledgments

I would like to express my deepest gratitude to all those who provided me with the opportunity to complete this report.

First and foremost, I am grateful to **Mr. Aymen Ayari**, my supervisor, for his invaluable guidance, encouragement, and insightful suggestions throughout the course of this project. His expertise and support were instrumental in shaping the direction and quality of this report.

I am also indebted to **The Wimbee Organization** for granting me access to their facilities and resources, which were crucial for the successful completion of this report. Special thanks to the staff members who assisted me in various capacities.

Additionally, I wish to thank my friends and family for their unwavering support and encouragement throughout this endeavor. Special thanks to my mom **Dalenda**, my dad **Khaled**, and my sisters **Senda** and **Rym** for their belief in me and their emotional support, which provided the strength and motivation needed to persevere.

Last but not least, I would like to express my appreciation to anyone who contributed, directly or indirectly, to the completion of this report. Your support and encouragement have been invaluable.

Thank you all.

**Ahmed Dammak**

# Approval

I certify that I am the author of this report and that any assistance I have received in its preparation is fully acknowledged and disclosed in this report. I have also cited any source from which I used data, ideas, or words, either quoted or paraphrased.

Further, this report meets all of the rules of quotation and referencing in use at TBS, as well as adheres to the fraud policies listed in the TBS honor code.

No portion of the work referred to in this report has been submitted in support of an application for another degree or qualification to this or any other university or academic institution.

# Abstract

”Without data, you’re just another person with an opinion.” - W. Edwards Deming

Data analytics involves the systematic computational analysis of data, allowing for the extraction of meaningful insights and patterns. This process leverages a variety of techniques, including statistical analysis, machine learning, and data mining, to interpret complex datasets and drive informed decision-making.

The impact of data analytics on people’s lives is profound and multifaceted. In the healthcare sector, it enables the prediction and management of diseases, improves patient outcomes, and optimizes operational efficiencies in hospitals. In education, data analytics helps in tailoring personalized learning experiences and identifying areas where students may need additional support. Retail businesses use data analytics to understand consumer behavior, forecast trends, and enhance customer experiences through targeted marketing strategies.

By converting raw data into actionable insights, data analytics not only drives business growth and innovation but also improves efficiency and quality of life. It is a testament to the power of information in shaping a more informed and responsive society.

Our project focuses on gathering, processing, and analyzing data related to your company to gain valuable insights and extract critical metrics. These insights will help us make informed decisions and take appropriate actions to improve various aspects of the business. To achieve this, our project begins with data gathering, where data from various sources within the company, such as Projects, Tasks, as well as others, is collected using both automated tools and manual methods. This involves integrating with existing databases and scraping data from relevant sources.

The collected data then undergoes the ETL (Extract, Transform, Load) process. During the extract phase, raw data is pulled from its original sources. In the transform phase, the data is cleansed and formatted to ensure consistency and accuracy, which includes handling missing values, removing duplicates, and standardizing data formats. Finally, in the load phase, the cleaned and transformed data is directed into a centralized data warehouse for further analysis. Following this, data visualization is performed to create visual representations of the processed data, making it easier to understand and interpret. Subsequently, in-depth analysis of the visualized data is conducted to extract meaningful insights, including trend analysis, predictive modeling, and identifying correlations. Key performance indicators (KPIs) and other relevant metrics are also identified.

# keywords

**BI:**Business Intelligence

**ETL:**Extract-Transform-Load

**CRISP-DM:**CRoss Industry Standard Process for Data Mining

**MYSQL:** My Structured Query Language

**ODS :** Operational Data Store

**DWH:** Data Warehouse

**DB:**Database

**KPI:**Key Performance Indicator

# Table of Contents

<b>1</b>	<b>General Framework of the Project</b>	<b>8</b>
1.1	Presentation of the Host Organization . . . . .	8
1.2	Improvement . . . . .	9
1.3	Project Context . . . . .	9
1.3.1	Analysis of Existing Solutions . . . . .	10
1.3.2	What Our Project Brings to the Table . . . . .	10
1.4	Requirements Capture . . . . .	11
1.4.1	Users Identification . . . . .	11
1.4.2	Functional Requirements . . . . .	11
1.4.3	Non-functional Requirements . . . . .	12
1.5	System Modeling . . . . .	13
1.5.1	Description . . . . .	13
1.6	Activity Domain & Tools Used . . . . .	14
1.6.1	Data Processing and Integration . . . . .	14
1.6.2	ODS and DWH Setup . . . . .	14
1.6.3	Visualization and Reporting . . . . .	15
1.6.4	Report Writing . . . . .	15
1.6.5	Project Planning . . . . .	16
1.7	Work Methodology . . . . .	16
1.7.1	Waterfall approach . . . . .	17
1.7.2	CRISP-DM approach . . . . .	18
1.7.3	Choice of approach . . . . .	19
<b>2</b>	<b>Technical Framework</b>	<b>20</b>
2.1	Introduction . . . . .	20
2.2	Business Intelligence Concept . . . . .	20
2.3	Data Mining . . . . .	20
2.4	ETL . . . . .	20
2.5	Databases (DB) . . . . .	21
2.5.1	Operational Data Store (ODS) . . . . .	22
2.5.2	Data Warehouse (DWH) . . . . .	22
2.5.3	Key Differences and Considerations . . . . .	22
2.6	Conclusion . . . . .	22
<b>3</b>	<b>BI Solution</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.2	Data Warehouse Architecture . . . . .	23
3.2.1	Top-down approach . . . . .	23

3.2.2	Bottom-up approach . . . . .	25
3.3	Data Warehouse Modeling . . . . .	25
3.3.1	Star Schema . . . . .	25
3.3.2	Snowflake Schema . . . . .	26
3.3.3	Constellation Schema . . . . .	26
3.4	Understanding the Data . . . . .	27
3.5	Physical Model Design . . . . .	27
3.6	Data Processing . . . . .	28
3.6.1	Creation of Databases . . . . .	28
3.6.2	Connection to Databases . . . . .	28
3.6.3	Creation of Tables . . . . .	29
3.7	Creation of the ETL . . . . .	30
3.7.1	Data Source . . . . .	30
3.7.2	ODS Loading . . . . .	31
3.7.3	DWH Loading . . . . .	31
3.7.4	Log Writing . . . . .	33
3.8	Reporting . . . . .	33
3.8.1	Connecting DB to Power BI . . . . .	33
3.8.2	Table Selection . . . . .	34
3.9	Dashboarding . . . . .	34
3.9.1	KPI Determination . . . . .	34
3.9.2	Dashboard Overview and results . . . . .	37

# List of Figures

1.1	Wimbee-Tech . . . . .	8
1.2	Proposed Solution . . . . .	10
1.3	Use Case Diagram . . . . .	13
1.4	Talend . . . . .	14
1.5	MySQL Workbench . . . . .	14
1.6	Power BI . . . . .	15
1.7	LaTeX . . . . .	15
1.8	Gantt Project . . . . .	16
1.9	Project Schedule . . . . .	16
1.10	Waterfall Methodology . . . . .	17
1.11	CRISP-DM Methodology . . . . .	18
2.1	ETL Process . . . . .	21
2.2	Database Setting . . . . .	21
3.1	Top-down Approach . . . . .	23
3.2	Bottom-up Approach . . . . .	25
3.3	Star Schema . . . . .	25
3.4	Snowflake Schema . . . . .	26
3.5	Constellation Schema . . . . .	26
3.6	Constellation Model . . . . .	27
3.7	Database Creation . . . . .	28
3.8	Database Connection . . . . .	29
3.9	ODS Tables . . . . .	29
3.10	DWH Tables . . . . .	30
3.11	Source Files . . . . .	30
3.12	ods_imputation loading . . . . .	31
3.13	ODS Insertion Mode . . . . .	31
3.14	DWH Loading . . . . .	32
3.15	Tmap Setting . . . . .	32
3.16	Logs Table . . . . .	33
3.17	Power BI - Database Connection . . . . .	33
3.18	DWH Table Selection . . . . .	34
3.19	Project Imputation . . . . .	37
3.20	Consultant Imputation . . . . .	39



# Chapter 1

## General Framework of the Project

### Introduction

To carry out this work properly, we will begin this first part of the report by presenting the host company which gave me the opportunity to improve my skills. Subsequently, we will present the project, the study of the existing situation, and the solution that we offer. We will then present the functional & non-functional requirements as well as the overall use case diagram. Finally, we will present the working methodology and the tools used which will help us to succeed in this project.

### 1.1 Presentation of the Host Organization



Figure 1.1: Wimbee-Tech

Wimbee-tech is a consulting firm specializing in data and digital strategy. Thanks to its 56 business experts, functional experts, technical developers, and consultants who master data innovations, digital, and market solutions. More than 120 achievements in the field of data and digital have enabled us to acquire extensive experience in project management.

Currently operating in France, Spain (Malaga), Poland, and Tunisia, Wimbee specializes in different areas:

- **Data:** The quantity of data within companies is growing, but this data is not well used. The importance of this area is to collect, connect, and analyze data to better exploit it and to have an optimal use.
- **Digital:** Digital transformation is imposed by the market, which is why Wimbee experts allow companies to adapt the right strategy that will help them to be competitive.
- **Consulting:** Based on a continuous improvement approach, Wimbee supports its clients on governance issues using techniques appropriate for each situation.
- **Systems Integration:** This area allows the company to interact and connect its different systems to be able to access reliable data. Wimbee supports its customers from the detection of the problem to the development and deployment of the solution.
- **Run:** Wimbee implements industrial strategies to support organizations for the development of new projects without forgetting to maintain their current systems. These strategies consist of mastering the production cycles and optimizing production costs.
- **Change Management:** During digital changes, the company must consider several factors that can fail this revolution.

## 1.2 Improvement

With a view to improvement & innovation, Wimbee is always looking for new collaborations with schools specializing in its fields of activity. It aims to apply new methods and technologies through its various research projects offered each year to students and future collaborators.

## 1.3 Project Context

In today's fast-paced business environment, decision-makers require timely and accurate insights to guide their strategic planning and operational decisions. Our project is designed to address these needs by providing comprehensive data analytics capabilities that allow for a detailed understanding of key business metrics. Specifically, decision-makers need insights on critical areas such as project completion rates, employee headcount, and projected turnover compared to previous years. These insights are essential for making informed decisions about resource allocation, strategic initiatives, and overall business direction.

### 1.3.1 Analysis of Existing Solutions

Currently, many organizations rely on fragmented systems and manual processes to gather and analyze data. These methods often result in several challenges:

- **Data Silos:** Data is stored in various isolated systems, making it difficult to get a unified view of the business.
- **Manual Processes:** Data collection and analysis are often done manually, which is time-consuming and prone to errors.
- **Lack of Real-Time Insights:** Existing solutions may not provide real-time data, leading to delays in decision-making.
- **Limited Analytical Capabilities:** Basic reporting tools may not offer advanced analytics such as predictive modeling or trend analysis.
- **Inconsistent Data Quality:** Data inconsistencies and inaccuracies can lead to unreliable insights.

### 1.3.2 What Our Project Brings to the Table

Our project addresses these shortcomings by implementing a robust data analytics framework that encompasses data gathering, ETL processes, and advanced visualization techniques. Here's how our project stands out:

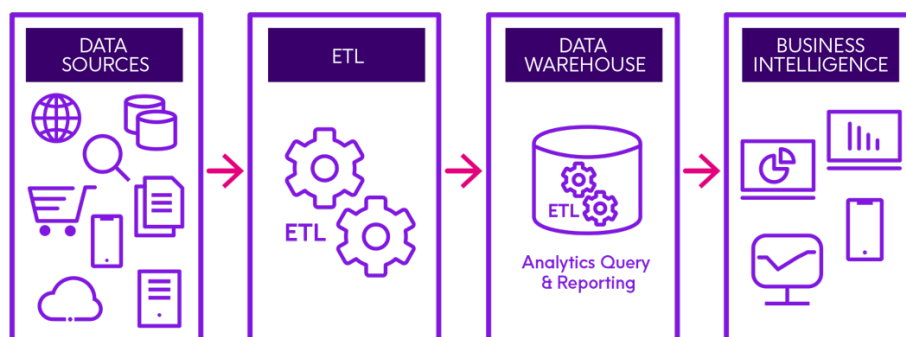


Figure 1.2: Proposed Solution

- **Integrated Data Collection:** By gathering data from various sources within the company, such as sales records, customer feedback, financial statements, and operational logs, our project ensures a comprehensive and integrated data collection process.
- **Efficient ETL Process:** The ETL (Extract, Transform, Load) process cleanses and transforms data to ensure consistency and accuracy. This involves handling missing values, removing duplicates, and standardizing data formats before loading the data into a centralized data warehouse.

- **Centralized Data Warehouse:** The cleaned and transformed data is stored in a centralized data warehouse, providing a single source of truth for all business data. This eliminates data silos and enables a unified view of the organization.
- **Advanced Data Visualization:** Utilizing tools like Tableau, Power BI, or custom dashboards, our project creates visual representations of the processed data. These visualizations highlight key metrics and trends, making it easier for decision-makers to understand and interpret the data.
- **Comprehensive Insights and Metrics:** The project enables in-depth analysis of the visualized data to extract meaningful insights. This includes trend analysis, predictive modeling, and identifying correlations. Decision-makers can track key performance indicators (KPIs) such as project completion rates, employee headcount, and turnover projections compared to previous years.
- **Data-Driven Decision Making:** With accurate and timely insights, decision-makers can make informed strategic decisions. This could involve optimizing marketing strategies, improving customer service, enhancing product offerings, or streamlining operations.
- **Scalability and Flexibility:** The project is designed to scale with the organization, accommodating growing data volumes and evolving business needs. It also provides flexibility in data analysis and reporting, allowing for customized dashboards and reports.

## 1.4 Requirements Capture

In this section, we identify the stakeholders of our model as well as the functional and non-functional requirements.

### 1.4.1 Users Identification

For the completion of our project, we have identified a primary user: the data analyst. The administrator's role is to collect and analyze data using an ETL process, and subsequently visualize it in order to help decision-makers take the appropriate action upon obtaining the results of the analysis.

### 1.4.2 Functional Requirements

Functional requirements are the actions that the model must perform and it only becomes operational if these are met.

In our case, our model needs to address the following requirements:

- **Process the data:** Understand and prepare the source data.
- **Create an ODS (Operational Data Store):** Establish a data layer between the source and the DWH (Data Warehouse).
- **Create a DWH (Data Warehouse):** Store historical data to support decision-making.
- **Visualize the data:** Generate visual representations to understand the results obtained.

### 1.4.3 Non-functional Requirements

Non-functional requirements are the criteria that judge the operation of a system, rather than specific behaviors. In our case, our model needs to meet the following non-functional requirements:

- **Performance:** The system should process and analyze data in a timely manner, ensuring minimal delay in generating recommendations.
- **Scalability:** The system must handle increasing volumes of data and more complex analyses as the dataset grows.
- **Reliability:** The system should consistently perform its intended functions without failure.
- **Usability:** The system should be user-friendly, enabling administrators to easily interact with it and understand its outputs.
- **Security:** Data must be protected against unauthorized access and breaches, ensuring confidentiality and integrity.
- **Maintainability:** The system should be easy to update and modify to incorporate new requirements or fix issues.

## 1.5 System Modeling

We will model our system's main functionalities as the following figure suggests :

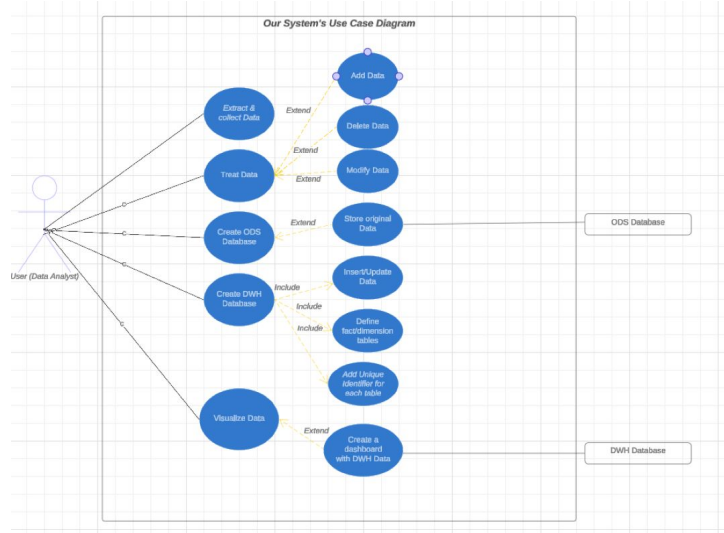


Figure 1.3: Use Case Diagram

### 1.5.1 Description

1. The user collects the data.
2. The user processes the data by adding, deleting, or modifying it.
3. The user creates an ODS (Operational Data Store) in which original data is stored.
4. Using the ODS data, the user defines fact tables and dimension tables, adds technical keys, inserts, and updates data to create the DWH (Data Warehouse).
5. Lastly, the user extracts the data from the DWH database and visualizes it using certain tools.

## 1.6 Activity Domain & Tools Used

For each step of the project, we utilized a variety of tools to ensure comprehensive and efficient execution as follows:

### 1.6.1 Data Processing and Integration



Figure 1.4: Talend

- TOS stands for Talend Open Studio for Data Integration.
- Talend is an ETL (Extract, Transform, Load) tool that extracts data from a source, transforms it, and then loads it into a destination.
- Data sources and targets can include databases, web services, or CSV files.

### 1.6.2 ODS and DWH Setup



Figure 1.5: MySQL Workbench

- MySQL Workbench is a visual tool designed for database architects and developers.
- It provides comprehensive administrative tools for data modeling, SQL development, server configuration, user management, and backups.
- It also offers connectivity with Talend to easily implement ODS and DWH.

### 1.6.3 Visualization and Reporting



Figure 1.6: Power BI

- Power BI Desktop is a free Microsoft application that installs on a local computer.
- It allows users to extract data from multiple sources, transform it, and create dashboards containing various visuals.

### 1.6.4 Report Writing



Figure 1.7: LaTeX

- LaTeX is a language and document preparation system.
- It consists of a set of macros designed to simplify text processing.
- The system automatically generates documents that adhere closely to typographical standards.



## 1.6.5 Project Planning

In a project, it is crucial to allocate tasks throughout the targeted period in chronological order.

This distribution helps identify each team member's responsibilities, the overall project duration, and the specific duration of each task. Effective planning ensures deadlines are met and provides a comprehensive overview of the project.



Figure 1.8: Gantt Project

We used GanttProject software, an open-source tool written in Java, to create a Gantt chart. You can view the chronological order of the project with estimated durations as follows :

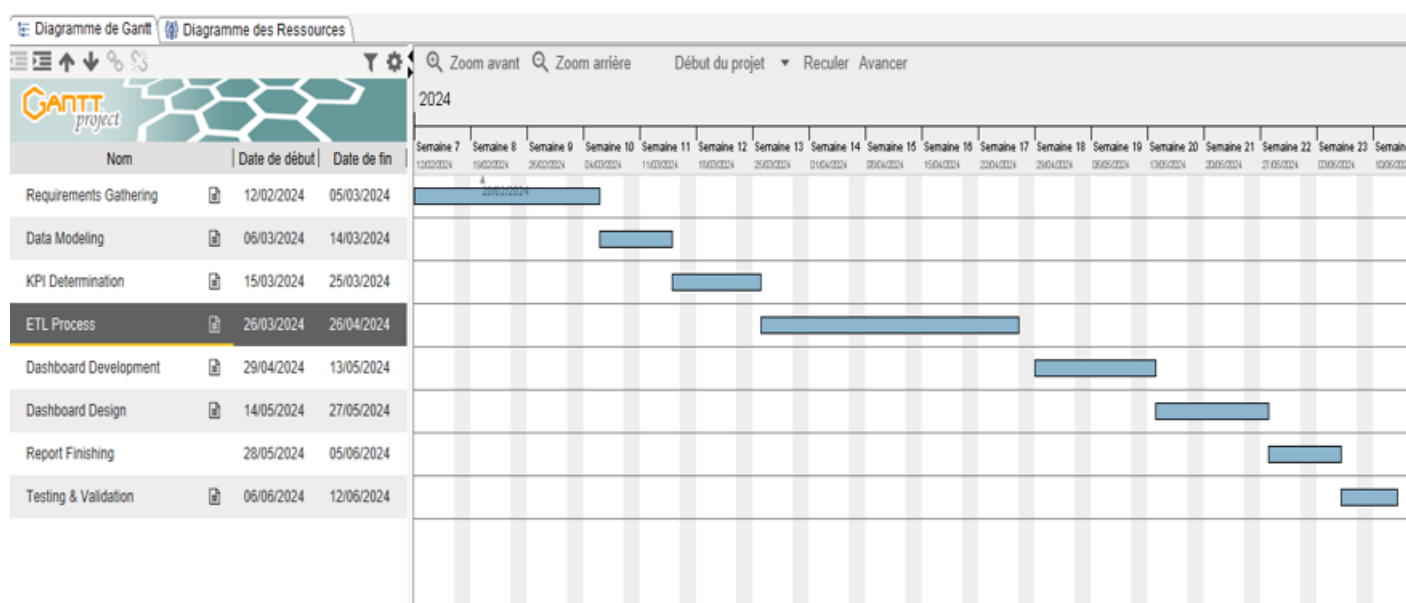


Figure 1.9: Project Schedule

## 1.7 Work Methodology

In order to carry out the work in an organized way , we will adapt to a certain work methodology that will allow to take matters on a step by step basis , therefore we will choose the most effective method from the following :

### 1.7.1 Waterfall approach

Waterfall methodology is a well-established project management workflow. Like a waterfall, each process phase cascades downward sequentially through five stages (requirements, design, implementation, verification, and maintenance) as shows the figure:

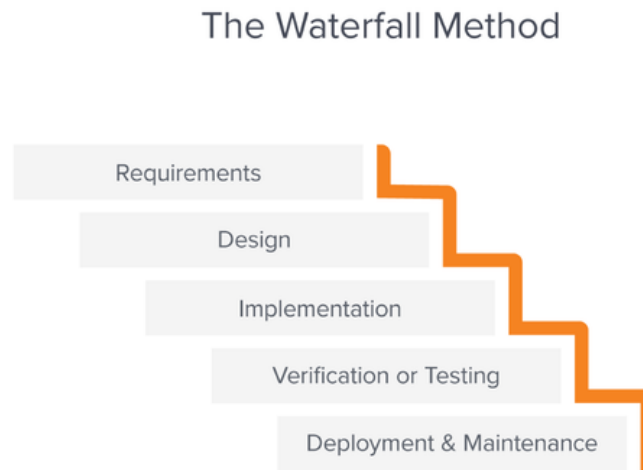


Figure 1.10: Waterfall Methodology

Unlike other methods, Waterfall doesn't allow flexibility. You must finish one phase before beginning the next. The project can't move forward until any problem is resolved. Moreover, if we're working as a team, we can't address bugs or technical debt if it's already moved on to the next project phase.

### 1.7.2 CRISP-DM approach

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It is a cyclical process that provides a structured approach to planning, organizing, and implementing a data mining project. The process consists of six major phases as shows the following figure:

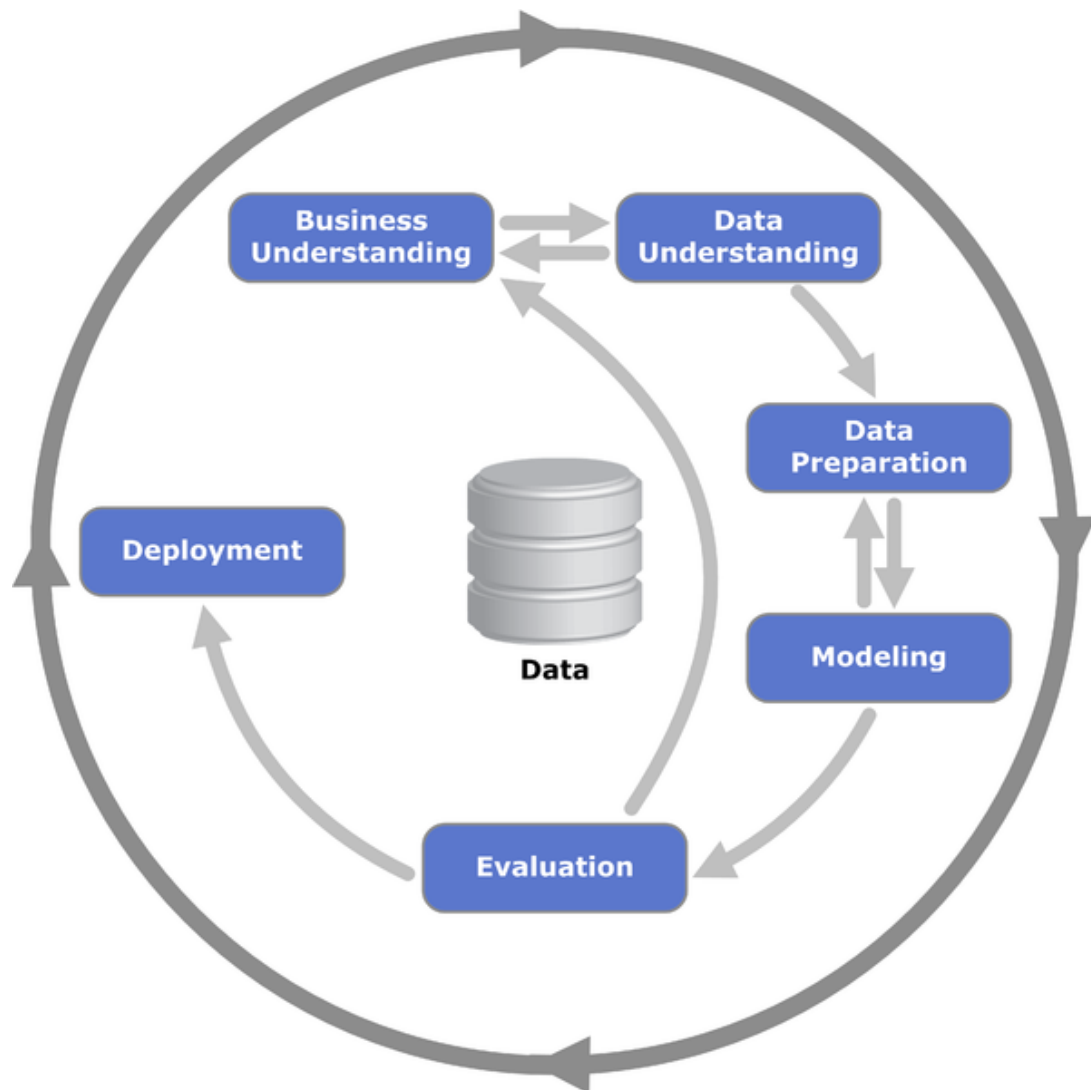


Figure 1.11: CRISP-DM Methodology

The CRISP-DM process offers a robust framework that guides data scientists and analysts in executing successful data mining projects. Its structured yet flexible approach ensures that all critical aspects of a project are addressed, from understanding the business problem to deploying the solution.

### 1.7.3 Choice of approach

We will compare each approach by weighing in on the pros and cons of each method in the following table:

Methodology	Pros	Cons
<b>Waterfall Methodology</b>	<ul style="list-style-type: none"><li>• Structured and easy to manage.</li><li>• Clear milestones.</li><li>• Extensive documentation.</li><li>• Predictable outcomes.</li></ul>	<ul style="list-style-type: none"><li>• Inflexible to changes.</li><li>• Late testing can reveal issues late.</li><li>• High risk if any phase has issues.</li><li>• Can be time-consuming.</li></ul>
<b>CRISP-DM Methodology</b>	<ul style="list-style-type: none"><li>• Tailored for data mining.</li><li>• Flexible.</li><li>• Iterative.</li><li>• Well-documented.</li><li>• Encourages high-quality, refined models.</li></ul>	<ul style="list-style-type: none"><li>• Requires significant up-front planning.</li><li>• Needs skilled personnel.</li><li>• Can be slow to identify data issues.</li><li>• May delay actual data mining work.</li></ul>

We can see that each approach has its own strengths and flaws. Nevertheless, our project revolves around Data Science which will bring us to adapt to the CRISP-DM methodology as it is the most appropriate one to adapt.

### Conclusion

During this chapter, we identified the functional and non-functional requirements of our project as well as showcasing it through the use case diagram, also we selected a variety of tools for the execution of our project along with adapting to a certain work methodology.

# Chapter 2

## Technical Framework

### 2.1 Introduction

In this chapter we discuss the different key concepts around our work and the different techniques that can be used in the context of this project.

### 2.2 Business Intelligence Concept

Business Intelligence (BI) is a process of collection, integration, analysis and communication of information allowing managers, supervisors and other end users in an organization to make business decisions illuminated.

Business intelligence includes several tools, applications and techniques that allow organizations to gather information from internal systems and from external sources. These data are then prepared for analysis in order to create reports and dashboards to provide analysis results to decision-makers.

### 2.3 Data Mining

The practice of analysing the big data present in datawarehouse is data mining. It is used to find the hidden patterns that are present in the database or in datawarehouse with the help of algorithm of data mining.

### 2.4 ETL

ETL is the abbreviation for Extract Transform Load. This is a type of software that collects data from various sources, converts it into a format suitable for a data warehouse and transfers them there.

More specifically, ETL works in three stages. In the extraction phase, the data is collected from one or more sources and is not under the same format. In the transformation phase, the data is reformatted and converted according to the needs. In the loading phase, the transformed data is transferred to the target data warehouse or database.

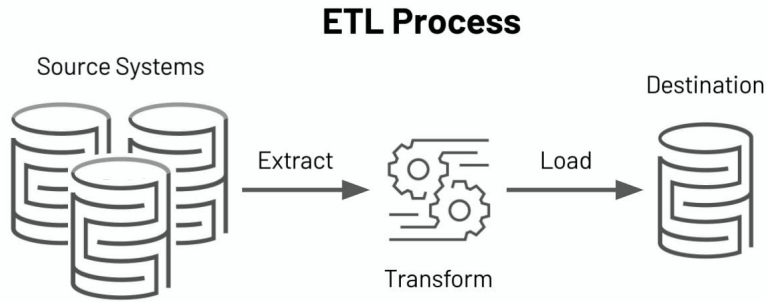


Figure 2.1: ETL Process

## 2.5 Databases (DB)

A database is a collection of information organized in such a way as to facilitate search, management and updating. In a database, the data are organized by rows, columns and tables. It is indexed for easy searching. Each time new information is added, the data is updated and, if necessary, deleted.

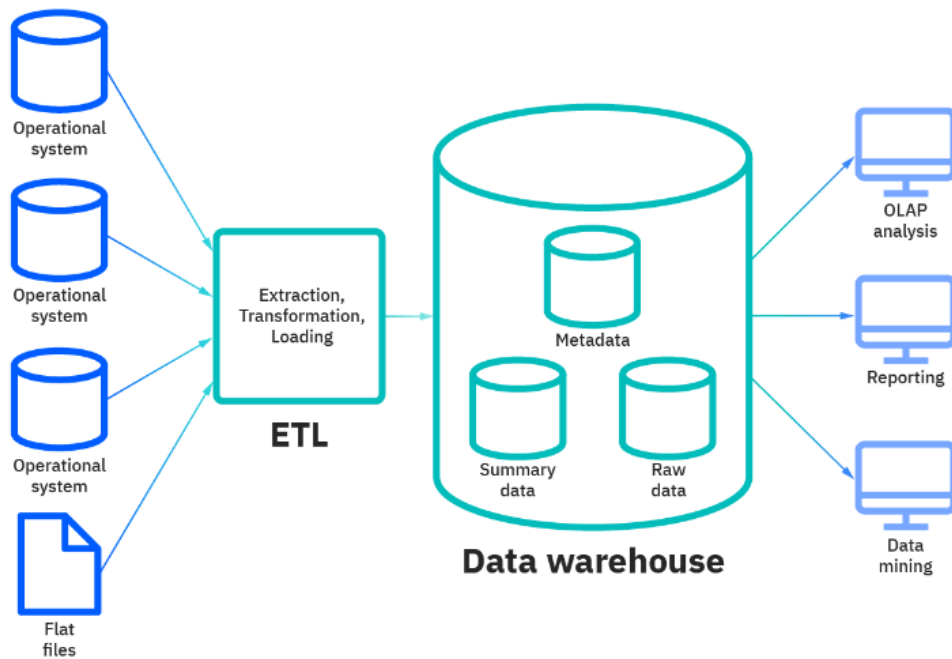


Figure 2.2: Database Setting

### 2.5.1 Operational Data Store (ODS)

Operational Data Store (ODS) is an intermediate structure for data management. It allows the source data of operational systems to be temporarily stored in with a view to their subsequent processing using specific tools. The data is extracted, filtered and collected in a second database. This database can be processed and additional information can be retrieved. This allows a faster access because redundant data is removed.

### 2.5.2 Data Warehouse (DWH)

Data Warehouse (DWH) is a relational database tool designed for data mining, analysis, decision making and Business Intelligence type activities, rather than for database applications such as transaction processing.

The information stored in a data warehouse is historical, structured, non-volatile and subject-oriented. These recordings provide details on various transactions over time. Data warehouses often contain data redundant to provide more information to users. That is why data stored in data warehouses is often aggregated so that the users can access it easily. In addition to relational databases, Data warehouse environments include ETL tools.

An essential characteristic of data warehouses is the organization of information by subject (customers, products, etc.). In reality, a data warehouse is defined by the type of data it contains and the people who use it.

### 2.5.3 Key Differences and Considerations

**Latency:** ODS often aims for lower latency compared to data warehouses. Real-time or near real-time insertion is more common in ODS, while data warehouses typically use batch or micro-batch ETL processes.

**Transformation:** In ODS, data is usually stored in a raw or lightly processed form to support operational reporting. In a data warehouse, data undergoes significant transformation to fit the analytical models and schemas.

**Data Volume:** Data warehouses handle large volumes of historical data optimized for complex queries, while ODS focuses on current or near-current data for operational use.

**Use Cases:** ODS supports operational tasks and real-time decision-making, while data warehouses are used for long-term trend analysis, business intelligence, and strategic decision-making.

## 2.6 Conclusion

The general idea of this chapter is to have a clear idea of the concepts that we are going to address throughout our project.

# Chapter 3

## BI Solution

### 3.1 Introduction

During this chapter, we will combine all the data coming from different sources to automate our work autonomously without human intervention. This part comes in different stages. First of all, we will start with the collection and understanding of data. Next, we will create our database as well as the tables which constitute it with SQL scripts on MySQL Workbench and the We will load data from multiple source files. We will create our ETL which contains several jobs on Talend.

There will be two stages of integration:

1. The integration of data from the ODS “Operational Data Store”: There will be all the transformation manipulations to be done on the source data.
2. The integration of data from the ODS to the DWH “Data Warehouse”: which represents our data warehouse from which we find the final version of our data.

We will also end this phase with the presentation of a dashboard in providing simple, consistent and relevant visual representations that allow us provide information to all stakeholders in an intuitive manner.

### 3.2 Data Warehouse Architecture

#### 3.2.1 Top-down approach

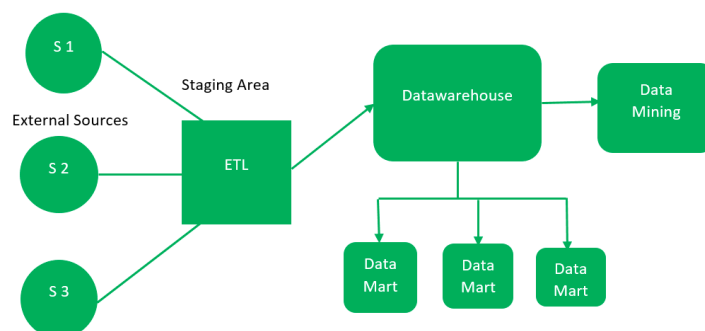


Figure 3.1: Top-down Approach



## External Sources

External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi-structured and unstructured as well.

## Stage Area

Since the data, extracted from the external sources does not follow a particular format, so there is a need to validate this data to load into datawarehouse. For this purpose, it is recommended to use ETL tool.

- **E (Extracted):** Data is extracted from External data source.
- **T (Transform):** Data is transformed into the standard format.
- **L (Load):** Data is loaded into datawarehouse after transforming it into the standard format.

## Data-warehouse

After cleansing of data, it is stored in the datawarehouse as central repository. It actually stores the meta data and the actual data gets stored in the data marts. Note that datawarehouse stores the data in its purest form in this top-down approach.

## Data Marts

Data mart is also a part of storage component. It stores the information of a particular function of an organization which is handled by single authority. There can be as many number of data marts in an organization depending upon the functions. We can also say that data mart contains subset of the data stored in datawarehouse.

### 3.2.2 Bottom-up approach

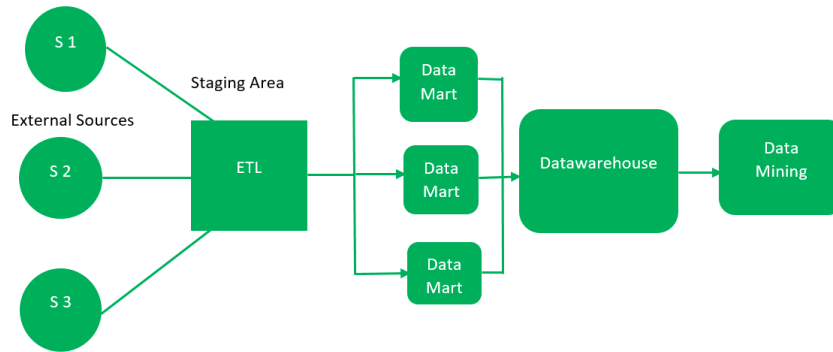


Figure 3.2: Bottom-up Approach

First, the data is extracted from external sources (same as happens in top-down approach). Then, the data go through the staging area (as explained above) and loaded into data marts instead of datawarehouse. The data marts are created first and provide reporting capability. It addresses a single business area. These data marts are then integrated into datawarehouse. This approach is given by Kinball as – data marts are created first and provides a thin view for analyses and datawarehouse is created after complete data marts have been created.

## 3.3 Data Warehouse Modeling

### 3.3.1 Star Schema

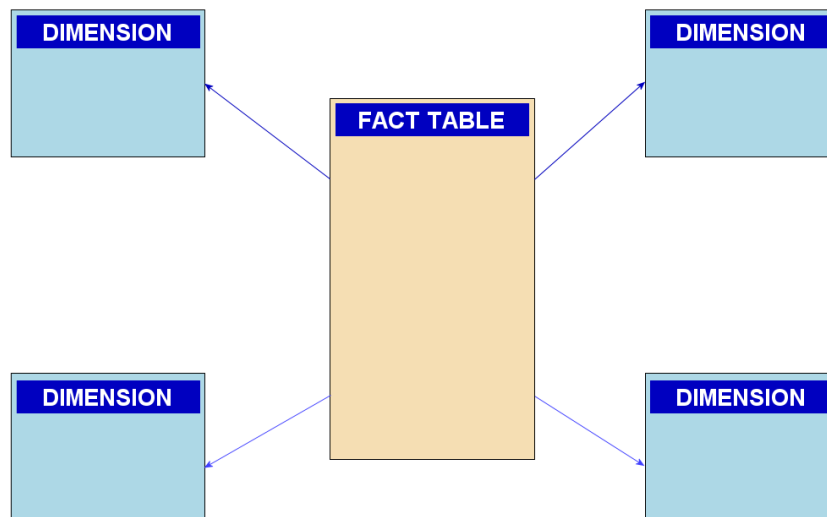


Figure 3.3: Star Schema

In a star schema, each dimension table is joined to the fact table through a foreign key relationship. This allows users to query the data in the fact table using attributes from the dimension tables.

For example, a user might want to see sales revenue by product category, or by region and time period. In our case, this schema can't be used to modelize our data.

### 3.3.2 Snowflake Schema

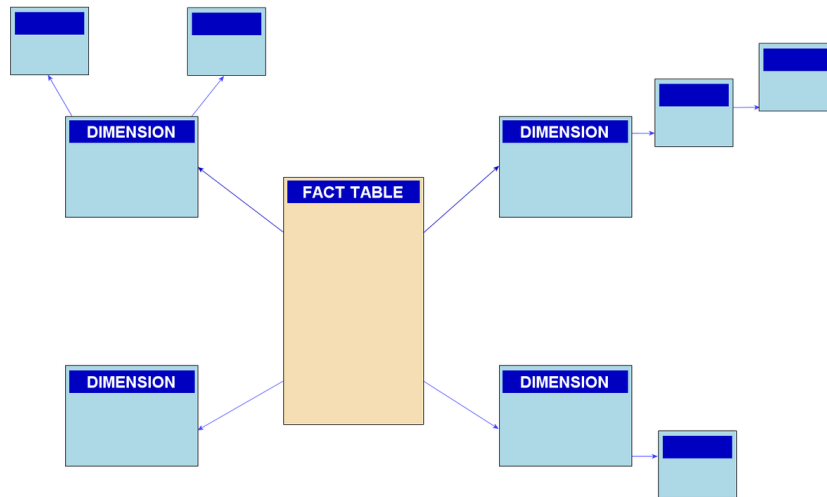
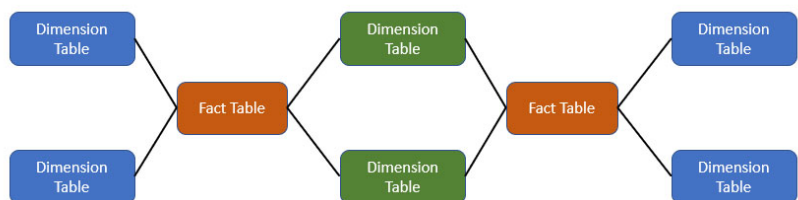


Figure 3.4: Snowflake Schema

It is a multi-dimensional data model that is an extension of a star schema, where dimension tables are broken down into subdimensions.

### 3.3.3 Constellation Schema



www.educba.com

Figure 3.5: Constellation Schema

In such model, there is more than a single fact table with tables of common dimensions. It can be considered as a multiplication of models in stars. This is one of the most used models in warehouse design of data.

## 3.4 Understanding the Data

In a BI project, it is essential to go through the data understanding stage. The first period of our internship was devoted to receiving data from the part of our supervisor, and subsequently the understanding of this data. This step allowed us to properly prepare the data provided and detect relationships between them. Our data is dedicated to the on-going projects within the organization. In fact it can be classified as follows :

- **Projects Information** : the statement of the basic details of a certain project(ID,Name,Description)
- **Projects Imputation** : the allocation of resources, time, and costs to specific tasks or activities within a given project.
- **Projects Capacity**: the acknowledgement of available resources, including personnel, equipment, budget, and time

## 3.5 Physical Model Design

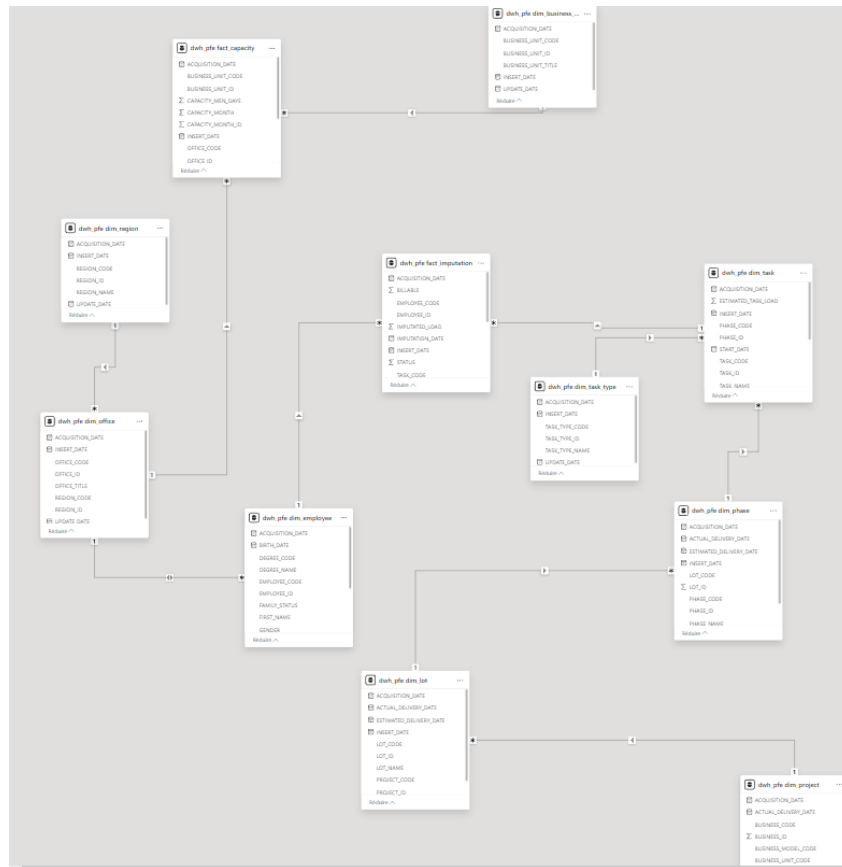


Figure 3.6: Constellation Model

For modeling our data, we designed a constellation model which presents the advantages cited without the previous part. This model must contain all the necessary data that can meet our needs. Our physical model of data is presented in the figure above.

## 3.6 Data Processing

Data processing makes it possible to extract information and knowledge from raw data. This process is usually automated using a computer.

The end product of this process will be presented to a human being, therefore the way it is presented is important for effective decision making. But the interpretation of the results varies from one person to another. So, in order to agree with all users of our solutions, it is essential for its ability to give meaning to the information, making it of better quality for use with other people, other processing or analysis tools. To do this, we often need to extract, transform and combine data to obtain information that everyone can understand and that meets the requirements of the final solution.

In this context, we began our treatment by creating the bases of data, subsequently the creation of tables and finally the feeding of these tables by the data. Moving from one stage to another requires data processing according to our need.

### 3.6.1 Creation of Databases

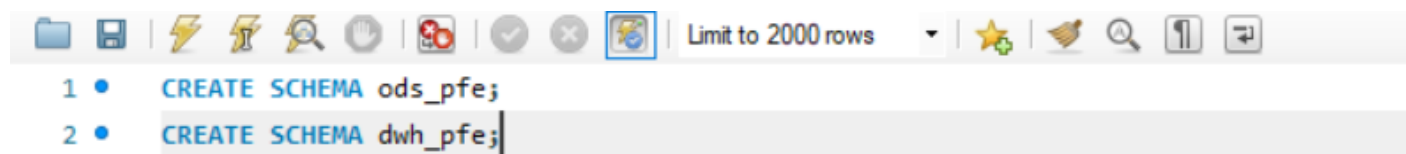


Figure 3.7: Database Creation

For the creation of the databases, we prepared two databases: one for the ODS tables and one for the DWH tables. We used the following SQL scripts to create these databases. Following the execution of these scripts we were able to obtain databases ready to welcome the different tables that compose them.

### 3.6.2 Connection to Databases

To be able to access the databases and create the tables, a connection must be configured to connect the ETL with the databases.

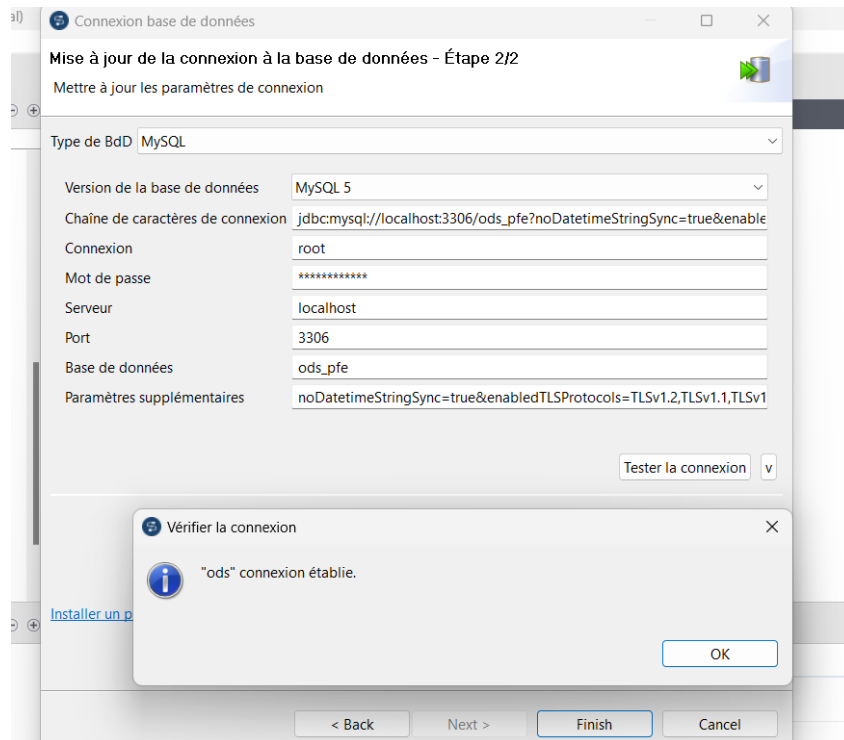


Figure 3.8: Database Connection

### 3.6.3 Creation of Tables

Once the connection is established, we can use Talend to create the different tables and display them in MySQL Workbench. There are source tables, ODS tables and DWH tables.

#### Creation of ODS Tables

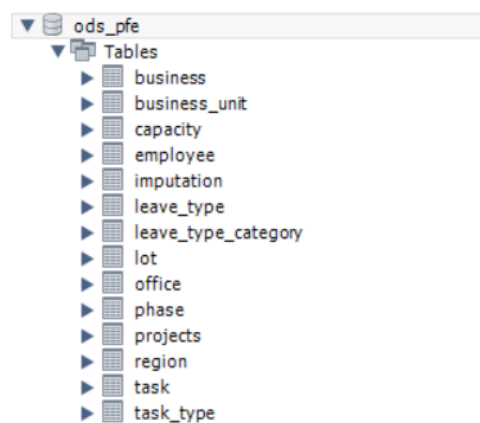


Figure 3.9: ODS Tables

ODS tables are used to store the different data sources that need to be processed before being sent to the DWH. Indeed, from the source tables we were able to create the ODS tables in the same way as the source tables.

## Creation of DWH Tables

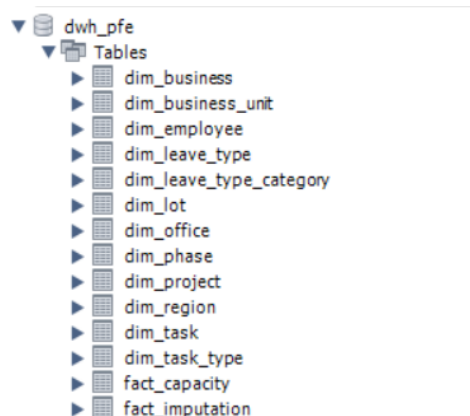


Figure 3.10: DWH Tables

The DWH tables represent our data warehouse. In fact, this is the final version data after ETL. The information obtained from these is then transferred to Power BI, our reporting tool. The feeding method used in the integration process is insert/update mode. However, a lookup is carried out always before each import to check if the imported data must be inserted or updated. The creation of the DWH tables is done from the ODS tables and will be injected into the database.

## 3.7 Creation of the ETL

After having created the different tables we will now move on to creating our ETL. He will be represented under several jobs. Each job represents a filled table subsequently in the database.

### 3.7.1 Data Source

Our data was extracted from excel files as presents the following figure:





Nom	Modifié le	type	taille
 FACT_OPS_IMPUTATION	05/03/2024 21:20	Feuille de calcul ...	81 Ko
 OPS_CAPACITY	05/03/2024 21:20	Feuille de calcul ...	21 Ko
 OPS_Projects	05/03/2024 21:19	Feuille de calcul ...	38 Ko
 REF_Data_Source	05/03/2024 21:19	Feuille de calcul ...	39 Ko

Figure 3.11: Source Files

### 3.7.2 ODS Loading

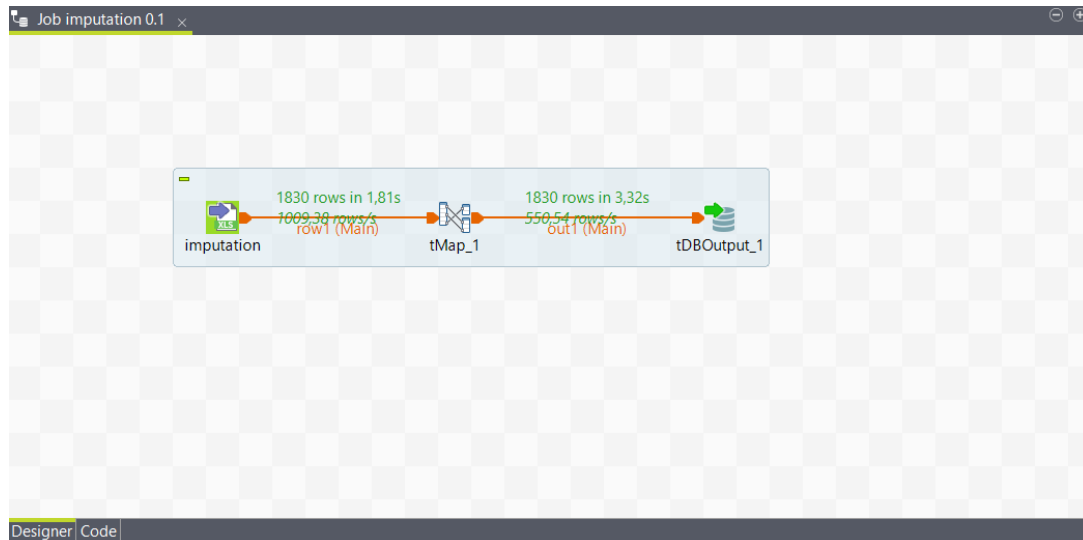


Figure 3.12: ods\_imputation loading

ODS (Operational Data Store) loading from Excel files involves importing data stored in Excel spreadsheets into an ODS, which serves as a centralized repository for real-time or near-real-time operational data. This process typically includes extracting data from Excel files, transforming it to meet the ODS schema requirements, and loading it into the ODS. the insertion mode of ODS data is defined as truncate/insert as following :

Figure 3.13: ODS Insertion Mode

### 3.7.3 DWH Loading

The DWH database consists of two parts: fact tables and dimension tables. Since fact tables are populated from dimension tables, we began by loading the dimension tables. For loading any table in the DWH, whether it is a dimension table or a fact table, we need its corresponding ODS table as input. This input allows us to perform the necessary transformations based on our requirements.



For example, consider the following dimension table "dim\_business," which uses the "ods\_business" table as its input. Subsequently, a lookup is performed using a tmap,

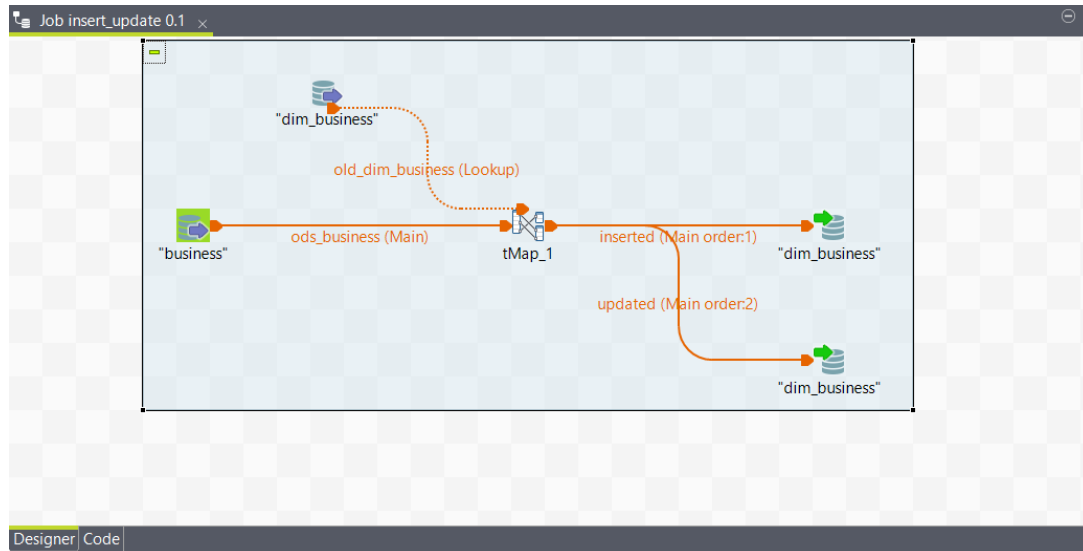


Figure 3.14: DWH Loading

which assists with an inner join to check whether the data from the ODS already exists in the DWH as we can see below in the figure:

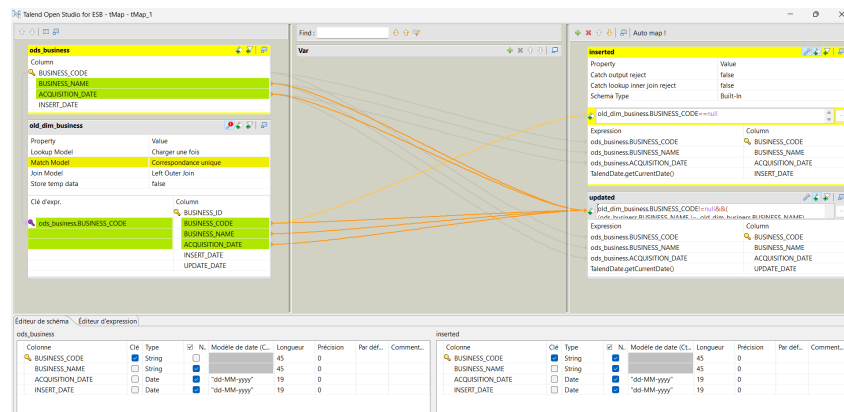


Figure 3.15: Tmap Setting

To achieve this, we created two outputs: "updated" and "inserted." The lookup checks if the input (ods\_business) does not exist in the output table (dwh\_dim\_business). If it doesn't exist, the action will be to insert it. If the input exists in the output table, the action will be to update it.

### 3.7.4 Log Writing

In order to keep up with data loading , we created a table by the name of 'logs' to record each job generation including its name,moment of creation,duration,status(success or failure) as well as source and destination files as shows the figure below:

project	job	pid	source	destination	moment	message	message_type	duration
PFE	ods_business	ZXMI7	OPS_PROJECTS.xlsx	ODS_PFE_BUSINESS	2024-06-05 14:16:45	success	end	3453
PFE	ods_business_unit	CuqN8	OPS_PROJECTS.xlsx	ODS_PFE_BUSINESS_UNIT	2024-06-05 14:17:36	success	begin	3402
PFE	ods_business_unit	CuqN8	OPS_PROJECTS.xlsx	ODS_PFE_BUSINESS_UNIT	2024-06-05 14:17:39	success	end	3402
PFE	ods_capacity	47UuHE	OPS_CAPACITY.xlsx	ODS_PFE_CAPACITY	2024-06-05 14:19:20	success	begin	3808
PFE	ods_capacity	47UuHE	OPS_CAPACITY.xlsx	ODS_PFE_CAPACITY	2024-06-05 14:19:24	success	end	3808
PFE	ods_employee	rEUXHr	EMPLOYEE.xlsx	ODS_PFE_EMPLOYEE	2024-06-05 14:20:21	success	begin	5899
PFE	ods_employee	rEUXHr	EMPLOYEE.xlsx	ODS_PFE_EMPLOYEE	2024-06-05 14:20:27	success	end	5899
PFE	ods_imputation	USX9p3	OPS_IMPUTATION.x...	ODS_PFE_IMPUTATION	2024-06-05 14:26:01	success	begin	3159
PFE	ods_imputation	USX9p3	OPS_IMPUTATION.x...	ODS_PFE_IMPUTATION	2024-06-05 14:26:04	success	end	3159
PFE	ods_leave_type_category	7hA3Q6	OPS_PROJECTS.xlsx	ODS_LEAVE_TYPE_CATEG...	2024-06-05 14:27:26	success	begin	3124
PFE	ods_leave_type_category	7hA3Q6	OPS_PROJECTS.xlsx	ODS_LEAVE_TYPE_CATEG...	2024-06-05 14:27:29	success	end	3124
PFE	ods_leave_type	TJ5gl	OPS_PROJECTS.xlsx	ODS_PFE_LEAVE_TYPE	2024-06-05 14:28:19	success	begin	2810
PFE	ods_leave_type	TJ5gl	OPS_PROJECTS.xlsx	ODS_PFE_LEAVE_TYPE	2024-06-05 14:28:22	success	end	2810
PFE	ods_lot	ORb2jB	OPS_PROJECTS.xlsx	ODS_PFE_LOT	2024-06-05 14:30:12	success	begin	3164
PFE	ods_lot	ORb2jB	OPS_PROJECTS.xlsx	ODS_PFE_LOT	2024-06-05 14:30:15	success	end	3164
PFE	ods_office	kMGES2	REF_DATA_SOURCE...	ODS_PFE_OFFICE	2024-06-05 14:31:16	success	begin	2729
PFE	ods_office	kMGES2	REF_DATA_SOURCE...	ODS_PFE_OFFICE	2024-06-05 14:31:18	success	end	2729

Figure 3.16: Logs Table

In summary, this phase allowed us to first create the databases, including ODS tables, and DWH tables, and then populate these tables with data by generating jobs for each table and log each operation in the meanwhile. The next step will be reporting, where all this information will be transformed into interpretable visuals to guide decision-making.

## 3.8 Reporting

The goal of this section is to visualize the data stored in our DWH by extracting the main KPIs and analyze them using clear visuals to empower decision making.

### 3.8.1 Connecting DB to Power BI

MySQL database

Server

localhost

Database

dwh\_pfe

Advanced options

OK

Cancel

Figure 3.17: Power BI - Database Connection

To import data stored in the database, it's necessary to establish a connection that allows us to collect all available information from the DWH tables, manipulate it, and transform it into various visuals. To do this, we access Power BI, select a MySQL database to obtain the data, and configure the connection using the database connection details, as shown in the figure above.

## 3.8.2 Table Selection

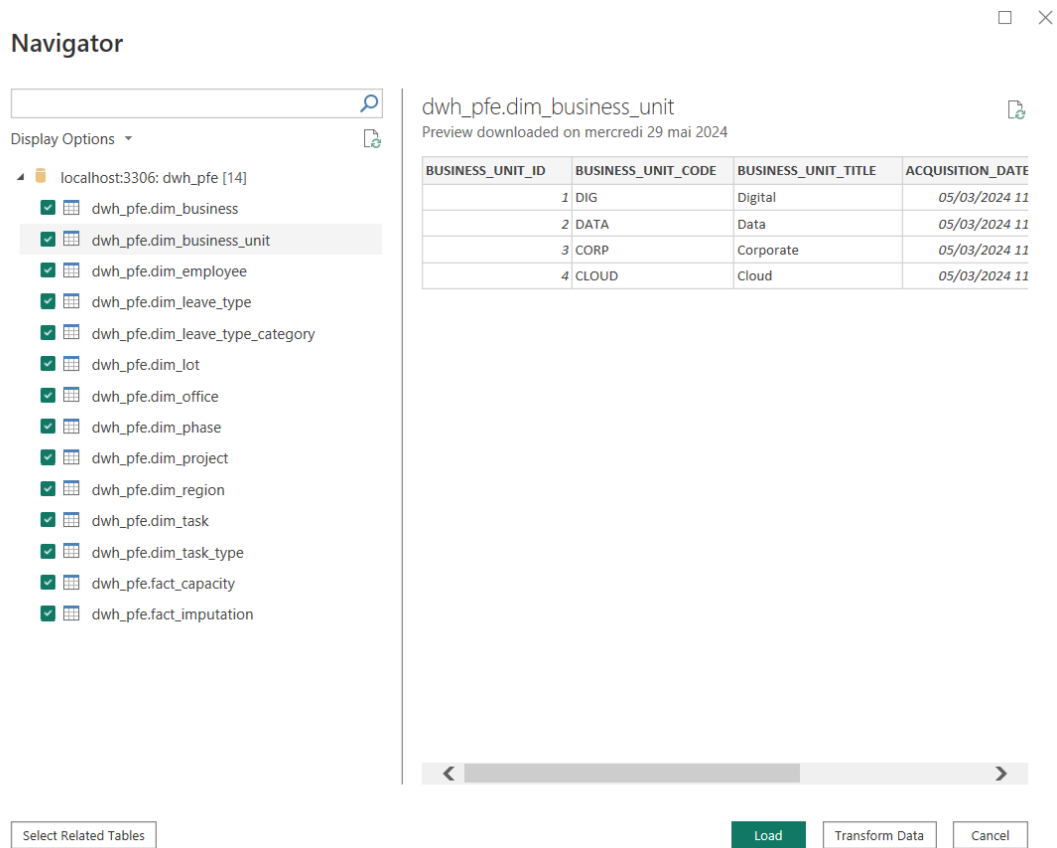


Figure 3.18: DWH Table Selection

Once we have access to the database, we will select all the DWH tables that will serve for the analysis phase. This connection will work as the bridge between MySQL and Power BI. Once new data is available in the database, we only need to go to Power BI and refresh the data to import the new records.

## 3.9 Dashboarding

This section is dedicated for building an operational dashboard that summarizes and contains various visuals, but first we need to determine the KPIs that will give us the needed insights on our data.

### 3.9.1 KPI Determination

We classified the dashboard into two main parts in which we will state all used KPIs as follows:

- **Project Imputation:** The resources allocated for any given project.
  1. Average Project Duration:
    - **Description:** The average amount of time taken to complete a project from initiation to closure.
    - **Visual:** Card
  2. Average Delivery Delay:
    - **Description:** The average amount of time by which a given project is delayed beyond its originally scheduled completion date.
    - **Visual:** Card
  3. Employee Count:
    - **Description:** The total number of employees currently working in the organization.
    - **Visual:** Card
  4. Project Count:
    - **Description:** The total number of projects currently active, completed, or initiated within a specific time period.
    - **Visual:** Card
  5. Total projects per year:
    - **Description:** The total count of projects that were started, finished, or ongoing within a particular calendar year.
    - **Visual:** Stacked Area Chart
  6. Total projects per office:
    - **Description:** The number of projects undertaken by each office within the organization.
    - **Visual:** Pie Chart
  7. Total projects per business unit:
    - **Description:** The amount of workload that each business unit is in charge of.
    - **Visual:** Clustered Bar Chart
  8. On-time delivery rate:
    - **Description:** The percentage of projects that are completed within the estimated delivery dates.
    - **Visual:** Gauge

- **Consultant Imputation:** The capacity assigned for any given task .
  1. Average Imputed Load per task:
    - **Description:** The average amount of load or effort assigned to each task in hours.
    - **Visual:** Card
  2. Non-billable load:
    - **Description:** The amount of work or effort that is not directly billable to a given project.
    - **Visual:** Card
  3. Tasks Count:
    - **Description:** The total number of tasks or activities that are being tracked, managed, or performed for a given project.
    - **Visual:** Card
  4. Total Capacity:
    - **Description:** The maximum amount of work days allocated for a set of projects.
    - **Visual:** Card
  5. Total Imputed Load:
    - **Description:** The maximum amount of work hours allocated for a set of tasks.
    - **Visual:** Card
  6. Monthly Imputed Load:
    - **Description:** The total monthly amount of work hours allocated for a set of tasks.
    - **Visual:** Line Chart
  7. Billable and non-billable load per year and month :
    - **Description:** The amount of billable and non-billable hours per year and month.
    - **Visual:** Clustered Bar Chart
  8. Billable load percentage:
    - **Description:** The percentage of billable hours for a set of projects.
    - **Visual:** Gauge
  9. Capacity per office:
    - **Description:** The capacity allocated in days for each office.
    - **Visual:** Pie Chart

### 3.9.2 Dashboard Overview and results

These are the different pages of the dashboard along with their respective insights :

- **Project Imputation:**

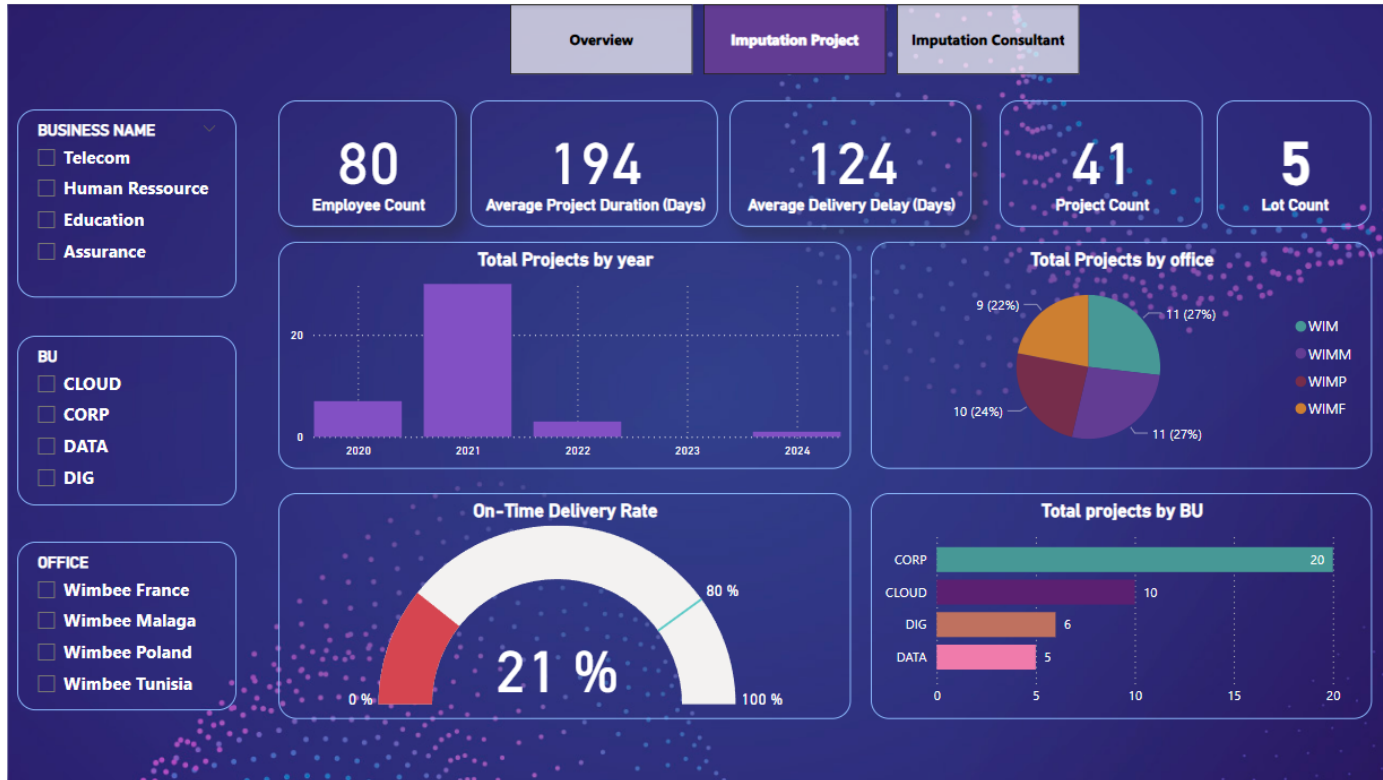


Figure 3.19: Project Imputation

#### Key Metrics

- **Employee Count:** 80
  - The total number of employees involved in the projects. This helps in understanding the workforce size relative to the workload and project demands.
- **Average Project Duration:** 194 days
  - This is the average length of time from project start to completion.
- **Average Delivery Delay:** 124 days
  - The average delay beyond the planned delivery date. High delays can signal issues in project management, resource allocation, or unexpected challenges.
- **Project Count:** 41
  - The total number of projects undertaken. This helps gauge the volume of work and resource distribution across multiple projects.

## Total Projects by Year

- **Trend (2020-2024):** The highest number of projects was in 2021, with a sharp decline in subsequent years.
  - This trend can help identify changes in project demand or strategic shifts in the organization's focus.
  - The decline in projects may require investigation into market conditions, internal capabilities, or changes in client needs.

## Total Projects by Office

- **Distribution:**
  - **WIM:** 11 projects (27%)
  - **WIMM:** 11 projects (27%)
  - **WIMP:** 10 projects (24%)
  - **WIMF:** 9 projects (22%)
- A relatively even distribution across offices indicates balanced project assignments and resource utilization.
- This helps in ensuring that no single office is overburdened or underutilized.

## Total Projects by BU (Business Unit)

- **CORP:** 20 projects
- **CLOUD:** 10 projects
- **DIG:** 6 projects
- **DATA:** 5 projects
- The distribution shows which business units are handling more projects, potentially indicating areas of higher demand or specialization.

## On-Time Delivery Rate

- **Current Value:** 21%
  - This indicates the proportion of projects delivered on or before the planned date. A low on-time delivery rate highlights potential inefficiencies and project management challenges.
  - Improving this rate is crucial for client satisfaction and maintaining a competitive edge.

- **Consultant Imputation:**

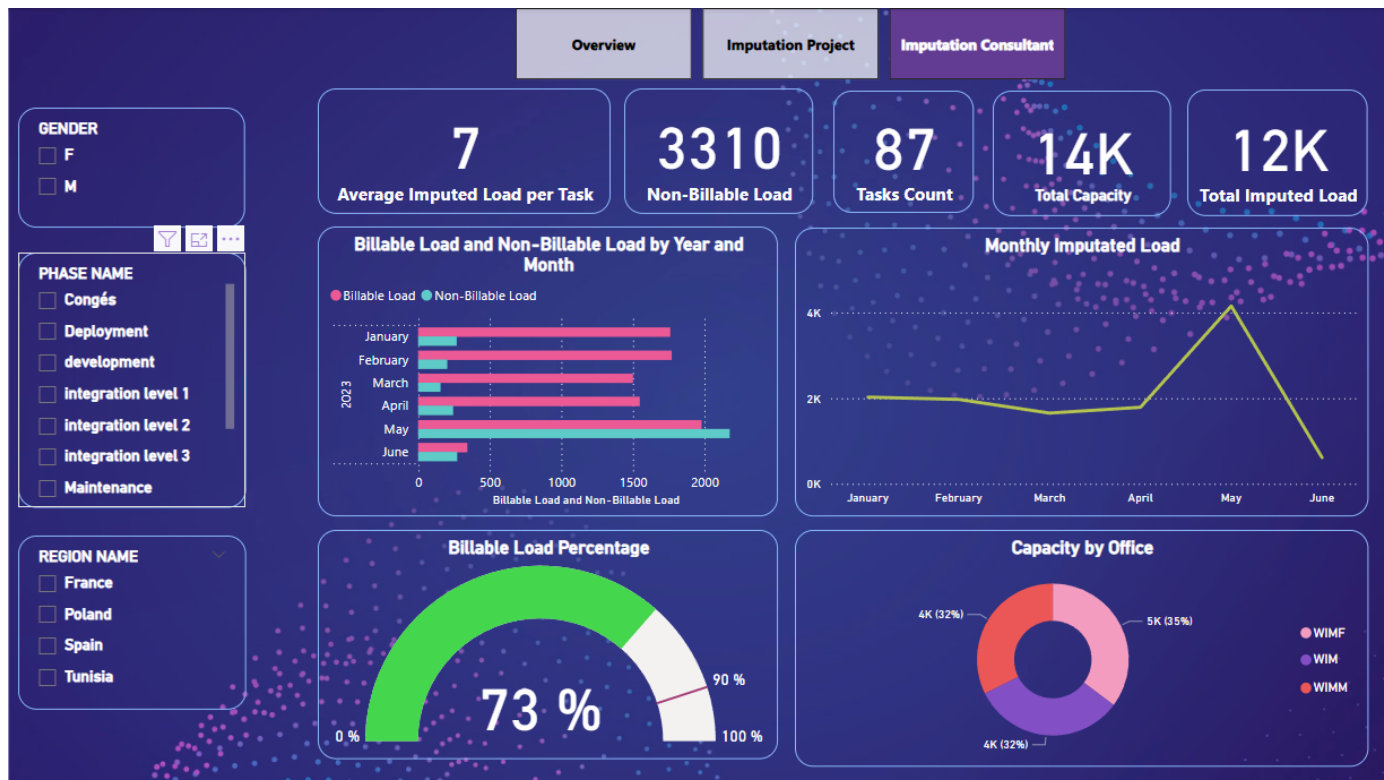


Figure 3.20: Consultant Imputation

## Key Metrics

- **Average Imputed Load per Task: 7**
  - This metric represents the average amount of workload imputed (assigned or logged) per task. A value of 7 suggests that on average, each task has a workload of 7 hours.
- **Non-Billable Load: 3310**
  - Non-billable load refers to the work that is performed but not directly charged to a client.
- **Tasks Count: 87**
  - The total number of tasks handled during the reporting period. Tracking the number of tasks can help in understanding workload distribution and productivity.
- **Total Capacity: 14K**
  - Total capacity represents the maximum potential workload that can be handled by the team or organization. It sets the upper limit for what can be accomplished within the given resources.



- **Total Imputed Load:** 12K
  - Total imputed load is the cumulative amount of work assigned across all tasks. Comparing this with total capacity helps gauge how much of the available capacity is being utilized.

## Billable Load vs Non-Billable Load (2023)

- **January to June:** The bar graph compares the billable and non-billable load for each month. Billable load is consistently higher than non-billable load which suggests more generation of revenue coming from the work charge to clients

## Monthly Imputed Load

- **Trend:** There is a peak in May, followed by a sharp decline in June.
  - Tracking monthly imputed load helps in identifying patterns and seasonal variations in workload.
  - The peak in May suggests a period of high activity, which might require additional resources or better planning in future similar periods.
  - The sharp decline in June could indicate project completions, reduced work, or efficiency improvements.

## Billable Load Percentage

- **Current Value:** 73%
  - This percentage represents the proportion of the total workload that is billable. A higher percentage indicates more revenue-generating work.
  - At 73%, this is a relatively healthy proportion, though there may still be room for improving billable activities.

## Capacity by Office

- **WIMF:** 5K (35%)
- **WIM:** 4K (32%)
- **WIMM:** 4K (32%)
  - This shows how the total capacity is equally distributed among different offices.
  - A relatively even distribution suggests balanced resource allocation across offices.

# Conclusion

This report was developed as part of our end-of-studies internship at Wimbee company, where we aimed to implement a Data Warehouse (DWH). To achieve our ultimate goal, we began by analyzing and situating our project within its context.

We identified both the functional and non-functional requirements, as well as the tools and technologies that assisted us in carrying out the various stages of this project. Subsequently, we proceeded through the necessary Extract, Transform, Load (ETL) steps for DWH implementation. The last section was dedicated to the analysis and visualization of our data in order to gain meaningful insights and help the company's decision makers to take action upon the results we got. Last but not least, This internship provided a highly enriching opportunity as it enabled us to consolidate and apply the academic knowledge acquired during our university studies. In addition to the technical aspect, we gained insight into the professional world and the challenges we may encounter.

# Bibliography

- [1] Talend. (n.d.). Retrieved from <https://www.talend.com/>
- [2] MySQL Workbench. (n.d.). Retrieved from <https://www.mysql.com/products/workbench/>
- [3] Power BI. (n.d.). Retrieved from <https://powerbi.microsoft.com/>
- [4] Excel. (n.d.). Retrieved from <https://www.microsoft.com/en-us/microsoft-365/excel>
- [5] LaTeX. (n.d.). Retrieved from <https://www.latex-project.org/>
- [6] Wikimedia Commons. (2023, June 10). CRISP-DM process diagram. Retrieved from [https://commons.wikimedia.org/wiki/File:CRISP-DM\\_Process\\_Diagram.png](https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png)
- [7] Atlassian. (n.d.). What is the Waterfall methodology? Retrieved from <https://www.atlassian.com/agile/project-management/waterfall-methodology> Data Warehouse Architecture. (n.d.). Retrieved from <https://www.geeksforgeeks.org/data-warehouse-architecture/>
- [8] ETL (Extract, Transform, Load). (n.d.). Retrieved from (Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.)
- [9] Operational Data Store (ODS). (n.d.). Retrieved from <https://www.techtarget.com/searchoracle/definition/operational-data-store>
- [10] Data Visualization. (n.d.). Retrieved from [https://www.tutorialspoint.com/data\\_warehouse/data\\_visualization.htm](https://www.tutorialspoint.com/data_warehouse/data_visualization.htm)
- [11] Data Analytics. (n.d.). Retrieved from [https://www.tutorialspoint.com/data\\_warehouse/data\\_analytics.htm](https://www.tutorialspoint.com/data_warehouse/data_analytics.htm)
- [12] Business Intelligence. (n.d.). Retrieved from [https://www.tutorialspoint.com/data\\_warehouse/business\\_intelligence.htm](https://www.tutorialspoint.com/data_warehouse/business_intelligence.htm)
- [13] Wimbee. (n.d.). Retrieved from <https://www.wimbee-tech.com/>
- [14] Tunis Business School. (n.d.). Retrieved from <https://tunis-business-school.tn/>
- [15] Big Data Framework. (n.d.). ETL in data engineering. Retrieved from <https://www.bigdataframework.org/knowledge/etl-in-data-engineering/>

- [16] TechTarget. (n.d.). Data collection. Retrieved from <https://www.techtarget.com/searchcio/definition/data-collection>
- [17] Software AG. (2023, April 4). Operational data stores (ODS) & data warehouses. Retrieved from [https://www.softwareag.com/en\\_corporate/blog/streamsets/operational-data-stores-ods-data-warehouses.html](https://www.softwareag.com/en_corporate/blog/streamsets/operational-data-stores-ods-data-warehouses.html)
- [18] Microsoft. (2023, June 10). Star schema. Retrieved from <https://learn.microsoft.com/en-us/power-bi/guidance/star-schema>
- [19] GeeksforGeeks. (n.d.). Functional vs non-functional requirements. Retrieved from <https://www.geeksforgeeks.org/functional-vs-non-functional-requirements/>
- [20] Wikipedia. (2023, June 1). Non-functional requirement. Retrieved from [https://en.wikipedia.org/wiki/Non-functional\\_requirement](https://en.wikipedia.org/wiki/Non-functional_requirement)
- [21] Wikipedia. (2023, March 5). Operational data store. Retrieved from [https://en.wikipedia.org/wiki/Operational\\_data\\_store](https://en.wikipedia.org/wiki/Operational_data_store)
- [22] RudderStack. (2023, April 22). The three stages of the ETL process. Retrieved from <https://www.rudderstack.com/learn/etl/three-stages-etl-process/>
- [23] GeeksforGeeks. (n.d.). Data warehouse architecture. Retrieved from <https://www.geeksforgeeks.org/data-warehouse-architecture/>
- [24] GeeksforGeeks. (n.d.). Star schema in data warehouse modeling. Retrieved from <https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/>
- [25] GeeksforGeeks. (n.d.). Snowflake schema in data warehouse model. Retrieved from <https://www.geeksforgeeks.org/snowflake-schema-in-data-warehouse-model/>
- [26] CRC. (n.d.). W. Edwards Deming: A short biography. Retrieved from [https://web.crc.losrios.edu/~larsenl/ExtraMaterials/WEDeming\\_shortbio\\_Ff4203.pdf](https://web.crc.losrios.edu/~larsenl/ExtraMaterials/WEDeming_shortbio_Ff4203.pdf)
- [27] Wiley. (2023). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). Retrieved from <https://www.wiley.com/en-gb/The+Data+Warehouse+Toolkit:+The+Definitive+Guide+to+Dimensional+Modeling,+3rd+Edition-p-9781118530801>