Dedication

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

It would be an honor to dedicate this modest work of graduation

To my dad Mohamed and my mom Lilia,

The two persons that gave the tools and value necessary to be where I am standing today. Nothing can express my respect, my eternal love, and my appreciation for all the sacrifices they have made for me, for their patience, love, and trust. I will never finish thanking both of them for all the opportunities that they have offered and given me, for all the teachings that they have told me, and for every piece of advice that came out of their mouth.

I hope that I can make them proud, the same way that I am proud of having both of them as my parents and as the compass of my life.

To all my family members, my sisters Mariem, Zineb, and Emna, my sister-in-law Nouha, my brother Hichem, and my brothers-in-law Mohamed & Ramzi,

As a token of my brotherly affection, my deep tenderness and gratitude, no dedication can express all the love I have for you.

To my nephews and nieces, Eya, Fatma, Mohamed, Kmar, Ali, Baya, Khadija & Soulaymen,

Thank you for giving my life its most beautiful colors with your joy and your cheerfulness.

To my sunshine,

Thank you for always believing in me, for helping and supporting me in achieving my dreams, and for bringing out the best in me by reminding me that I am always enough and I can do it.

You taught me to take risks and pursue my dreams with passion and integrity.

To all my friends and beloved ones,
Thank you all for the love, the encouragement, the beautiful memories, and the crazy moments.

To the people who enjoy reading these lines.

Appreciation

I would like to express my gratitude and my thanks to my supervisor Mrs. Mariem BEN HASSEN for the confidence that she granted me by agreeing to direct my work of memory. It is in fact thanks to her invaluable help, her relevant comments, and her continuous encouragement, that she never stopped giving me that this manuscript could see the day. I will always be grateful to you.

I sincerely thank Mrs. Raida KTARI, for the honor she gave me by accepting to be the president of my Jury.

I would also like to thank Mrs. Imen TOUNSI for accepting to review my memory.

Special thanks go to Mrs. Imen MOALLA, Haytham GHARAM, Ghanmi KHIR, Ayoub KALLEL, Mariem BEN SLAMA, Youssef BOUZGENDA, Mohamed Mseddi, Mariem Souayah, and Karim AMMAR for supporting me and helping me through the difficult times.

I do not forget to express my feelings of gratitude to all those who have contributed in any way to the success of this work.

I also thank all ISIMS administrative staff and teachers for the quality training and the atmosphere they have established during all our years of studying.

Finally, I would like to express our friendship and deep respect to all of my ISIMS colleagues.

Table of content

Table of figures

Table of tables

General Introduction

General Introduction

Decision-making has become an increasingly important task in today's highly competitive business world. As a result, many businesses are collecting large amounts of data from traditional information systems and making it available to decision-makers via decision support tools. These tools are known as Data warehouses. Of course, the presence of relevant information influences the quality of decisions. This information, however, does not come solely from internal company sources, but also from external sources.

Indeed, with the emergence of the Internet and social networks, these external sources have significantly increased the volume of information available, which is characterized by highly heterogeneous data.

On the other hand, many businesses are interested in analyzing and comprehending such social networks in order to assist decision-makers in making decisions. This has prompted several researchers to invest in the topic of analyzing social network opinions in order to improve decision-making and optimize results.

In the absence of appropriate tools, the analysis of business activity becomes a tedious task for the decision-makers of the "Re-searchlight" company. Furthermore, complex requests were required, which were costly in terms of response time and data-processing resources. As a result, the company must implement a decision-making system capable of meeting these requirements, as is the case with Business Intelligence (BI) tools. This will enable the collection of information required for the installation of dashboards tailored to the needs of decision-makers. Through BI applications, the company will be able to gain a perfect image of its activity and facilitate the taking of rigorous decisions.

The theme of our dissertation "Analysis of user profiling and behavioral data in the medical sector" is situated within this framework. The main objective of this project is to analyze the members' profiles of target Facebook groups related to the health field, based on their connection data and web-behavior.

Concretely, this project involves the development of a dashboard that collects large amounts of data from various Facebook groups (Big Data), intending to assist the business manager in his decision-making through BI technology in order to improve profitability and accelerate production.

This report is organized into six chapters. We present the overall context of our project in the first two chapters. In the remaining chapters, we explore our contribution.

Chapter 1- General scope. Firstly, this chapter will include a brief presentation of the general context of this project. In particular, we will present the host organization, emphasizing the analysis of the existing and its criticism in order to define the objectives to reach as well as the proposed solutions. Secondly, it will present the adopted project development methodology.

Chapter 2- State of the art and preliminary study. The purpose of this chapter is to present the key concepts of our field of study. Firstly, we will introduce the BI domain by discussing the concepts of data warehouse, ETL, etc. Thereafter, we will present the Open-Source Intelligence (OSINT) as the proposed solution. Finally, we will briefly address the notion of the user profile.

Chapter 3-Sprint 0-Sprint Backlog: Specification of needs and elaboration of the product backlog. This chapter will represent the first step in the Scrum methodology process, namely the "Sprint Zero," which specifies, on the one hand, the functional aspect of our application (i.e., the identification of functional requirements, the creation of the product backlog and the planning of our project's sprints) and on, the other hand, the technical study of our hardware and software environment.

The next three chapters entitled **Sprint 1- Data extraction**, **Sprint 2-Data transformation, and loading**, and **Sprint 3- Data visualization** will present in detail the Data Warehouse construction approach of our BI application. In each chapter/sprint, we will present a sprint backlog, a conception model, a realization step, and a sprint review.

We conclude this report with a general conclusion of all the work performed and open up potential opportunities for our application.

CHAPTER 1
GENERAL SCOOP

Chapter 1
General Scope

1.1. Introduction

In this chapter, we first present the host organization in which our project took place. Then, we present the general context of this project emphasizing the analysis of the existing and its criticism in order to define the objectives to reach as well as the proposed solutions. Finally, we define the development and project management methodology and the conceptual model that we have adopted.

1.2. Company presentation

1.2.1. Company description

Re-searchlight is a company located in Tunisia that specializes in the health field. Since its creation in 2007, it aims at accompanying and supporting the pharmaceutical sector. Its mission consists in providing their clients with key strategies about patients and principal prescribers.

This company works with more than 1000 doctors and has over 80000 network members. It has more than 14 years of expertise in several therapeutic sectors, like rare diseases, diabetes, and more.

1.2.2. Company services

Re-searchlight provides several services:

- **Patients support programs**: Re-searchlight manages patient support programs from conception to implementation.
- **Campaign management**: Re-searchlight sets up different types of campaigns for its clients. Awareness, communication, information, or explanation, Re-searchlight's teams accompany and advise its clients to meet their needs by proposing the best methodology appropriate to the target, the duration, and the budget.

- **Pharmacovigilance management**: Re-searchlight follows and adapts to the processes of multinationals and countries in terms of pharmacovigilance claims. Its team implements a centralized monitoring system allowing for accurate reporting while respecting deadlines imposed by clients
- **Market studies**: Re-searchlight has conducted hundreds of market research studies for its clients in the pharmaceutical industry. It provides support from conception to implementation, including an in-depth analysis of the data collected.
- **Digital services**: Re-searchlight offers a range of digital services to meet many needs. It brings its expertise in several fields: management of social networks, website development, content creation, video design and reporting, event management, and more.

1.2.3. Contact

- Company manager: Amel Bourassi
- Address: 4. St. Salah Dey, Menzah 4, Tunis
- Phone: +216 24 873 846
- Fax: +216 71 754 368
- Email: contact@re-searchlight.tn

1.3. Project analysis

1.3.1. Study of the existing and Problematic

Re-searchlight has many clients and over 86 000 followers on social media. Its database only contains information about clients they worked with, either inside or outside Tunisia.

As the number of people interested in the medical field grows day by day, Re-searchlight does not have a system that collects the large amount of users' information in its database. Thus, the company wants to save information about its followers on social media as well as users who are interested in the medical field. Moreover, it wants to get continuous reports about the collected information in order to enrich its database, make rigorous business decisions, and satisfy its business needs.

To do that, the company needs a system that extracts the followers' data, treats and analyzes them, as well as extracts useful information in order to get actionable insights and have real-time data at all times. Therefore, our mission is to make in place an information system specific to decisional applications that offers online data analysis to fulfill the needs of decision-makers. The decisional application must give quick and easy access to strategic information, and give actionable insights.

1.3.2. Proposed solution

As we mentioned in the previous paragraphs, a good data collection and analysis system is necessary to facilitate the making of good and rigorous business decisions.

The objective of our mission is to find a solution ensuring the link between the different data from different sources in order to achieve a restitution of information, which allows generating reports according to the needs.

To do this, we need a database dedicated to multi-dimensional analysis, and allows a link between existing information. This is where the interest in business intelligence comes from in this context. Also, our solution must not only meet the needs of analysis and processing of data, but it must satisfy the requirement of the huge volume of data that we will process, and gives results quickly, and this is where the interest of Big data comes from. Moreover, our system must collect data from free, public, and open sources (i.e., social networking sites (Facebook, Twitter, etc.)). In addition, this is where the OSINT interest comes from.

In conclusion, a proper combination of the concepts of big data, OSINT, and Business intelligence considerably improves the efficiency of the overall organization. In fact, Big Data will make BI a more valuable and useful tool for our project. Thus, Big Data techniques will improve our BI process by reducing response time and ensuring fast calculation.

1.4. Adopted development methodology

The project management methodology has a significant influence on the success of an IT project. In order to be successful in our project while respecting costs, deadlines, and objectives, it is necessary to be well organized and efficient. In this context, we had to set one of the existing project management methods to help us organize our project in a streamlined and structured way.

Choosing the methodology well leads to allowing all the team members to work with ease together while following defined rules.

1.4.1. Software development process models

When starting a new application or software development project, it is important to consider the various steps required for its final deployment, vis. analysis/specification, design/conception, implementation/coding, testing, and regular maintenance.

There are several types of software development and lifecycle management models. We mainly distinguish: (i) incremental models (e.g., Waterfall Model, V-Model) and (ii) non-linear models (e.g., RAD Model, Incremental Model, Iterative Model, Spiral Model, Prototype Model and Agile Model). The following table presents the advantages and disadvantages of each model.

Table 1.1 Software development model advantages and disadvantages

Model	Advantages	Disadvantages
Waterfall	- Ideal for lifecycle management of small projects where requirements are established and finalized upfront. - It is simple and understandable.	- Rigid structure. - Does not work well for complex projects where there is a possibility of requirement changes.
V	- Great for small projects. - It has test plans, and regular schedule updates throughout its lifecycle.	- Rigid structure - It is not ideal for applications or systems software that may require unforeseen changes/updates throughout the software lifecycle.
Incremental	- Great for projects who need some changes requests between increments. - It gives the ability to detect problems earlier in the software development for better lifecycle management planning.	- It tends to require more resources, staff and monetary, behind the project. - It is not ideal for ongoing development, as the next sequence cannot begin until the previous stage has fully completed.
RAD	- It reduces the development time. - It allows customer feedback throughout the software development.	- It requires highly versed developers as well as excellent modeling and planning skills.

		- Issues with final component assembly could result in unanticipated setbacks and the redevelopment of components to properly fit the rest.
Iterative	- It is easy to identify problems early when using this software development model.	- This model can take longer and be costly because of its rigid and no-overlaps phases.
Spiral	- It manages risks and divides development into phases. - Roadblocks are discovered earlier - It contributes to more accurate budget and schedule estimates.	- It also requires team members that are well versed in risk evaluation. - This model is so customizable, and repurposing the process can be confusing.
Prototype	- It reduces time and costs - User involved	- It can cause user confusion between prototype and finished product. - It can potentially add excessive development time for prototype development.
Agile	- It decreases the amount of time to yield individual system features. - It requires extensive communication and continuous feedback from the customer/user, which can provide clear direction for the project.	- An agile software development strategy has a minimal of documentation and requires a well-versed, cross-functional team. - It relies on end-user interaction that may or may not be clearly expressed.

In addition, Table 1.2 provides a further comparison between the predictive/traditional approach (e.g., waterfall, V) and the adaptive approach (agile method). Based on our previous analysis, we note that predictive methods are too rigid or not flexible enough, not agile enough. In contrast, agility is a rising value in the software production industry. Agility requires adapting to the change inherent in turbulent environments. These changes can be in specifications, stakeholders, procedures, and especially, in the software domain, in technologies.

Table 1.2 Predictive approach Vs Adaptive Approach

Predictive approach	Adaptive approach
Clear objectives through time	Short duration iteration with high quality increment
Process that seems simple and easy will lead to customer dissatisfaction (80% of software functions are little or not used)	Strong customer involvement
Saving time and money through high risk	Less risk
Planning-driven	Value-driven
Waterfall/V	Agile

From the list of selected software development models, agile model/approach seems to be the most promising and appropriate method for our project management and development.

1.4.2. Agile development method

1.4.2.1. Definition and characteristics

Agile methodology fosters an environment of adaptation (i.e., which is based on iterative and adaptive development cycles according to the evolving needs of the client), teamwork, self-organization, and rapid delivery which allows for a high level of client participation early in the project planning process. [1]

Agile methods require the involvement of both incremental and iterative models, as well as a high degree of adaptability:

- Gradual modification of software goals and specifications,
- Delivery of software that works immediately with a limited set of functionalities, where traditional methods require that the software be delivered only at the end of the complete development. This set of functionalities is then enriched regularly, according to new user requests or competitive pressure.
- The ultimate goal is user satisfaction in terms of perceived value, ease of use, and timeliness.

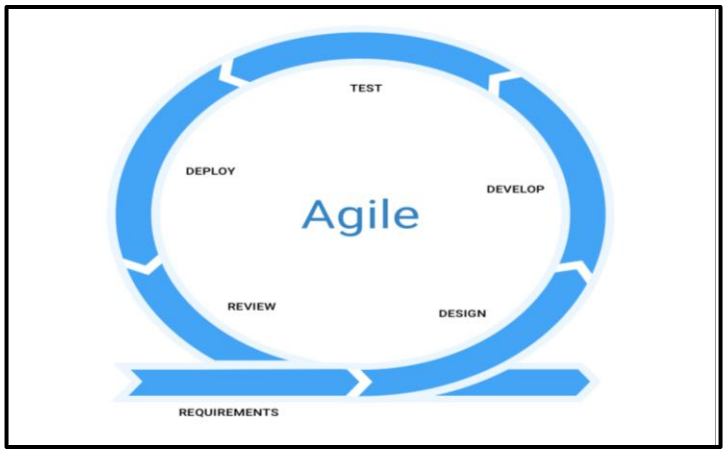


Figure 1.1 Agile Model

1.4.2.2. Benefits of using Agile

In this project, we take advantage of the agile method, which is summarized in the following points: [1]

- Always focused on the final product: a common vision for the team (customer satisfaction), the project is divided differently (by functionality) and organized in reduced development cycles (iterations).
- Collaboration with the client (avoiding the perverse effects of a contract).
- Adaptable: reactive to new needs and receptive to new solutions.
- Quality control increased: a step will not be validated if a deliverable does not suit the client, which avoids the accumulation of errors throughout the project.
- Superior quality product, better control, improved project predictability, reduced risks, increased flexibility, continuous improvement, etc.

1.4.2.3. Agile frameworks

More than a dozen methods claim to be part of the agile current: Rational Unified Process (RUP), Scrum, Kanban, eXtreme Programming (XP), etc. (Cf. Figure 1.2).

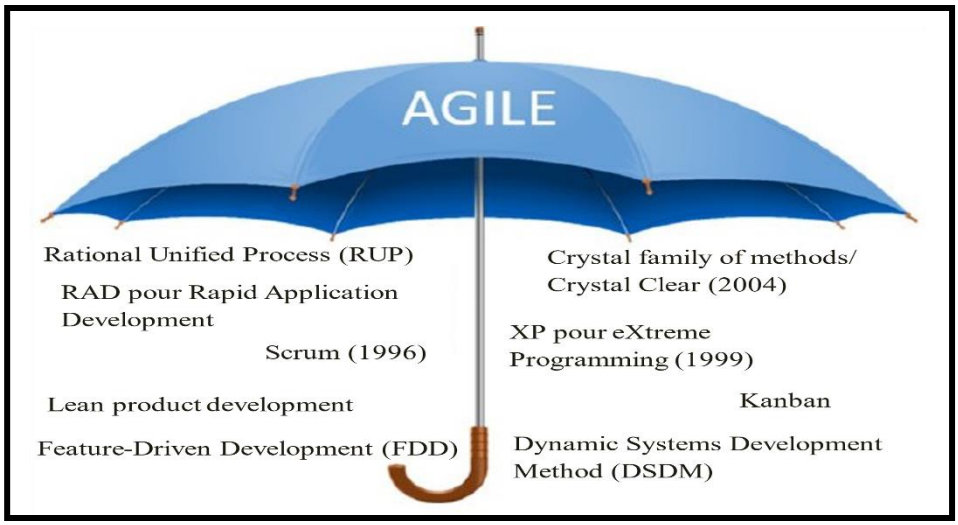


Figure 1.2 Agile Umbrella

Table 1.3 presents an extract of a comparative study of the different agile frameworks, in order to choose the most appropriate for the development of our project.

As a result of this comparative analysis, we will adopt the scrum framework. In fact, it represents an iterative, incremental, empirical approach and “Lean” thinking to optimize predictability and control risks (produce maximum value for minimum cost). Moreover, it offers a framework for working in multidisciplinary and self-organization teams, which are integrated and welded in a flexible, light and simple to understand strategy. It relies on self-managed and adaptable teams and on collective intelligence as well as a short development cycles "Sprints" (1-4 weeks). The ability to incorporate changes as they occur into a project currently in progress leads to better user satisfaction, better employee morale, higher quality, and higher productivity. In addition, scrum prioritizes the tasks by order of importance and this usually means that tasks to be completed first will probably affect the return on investment the most. Finally, it is simple to understand. It is the most used among these frameworks, the most proven, documented, and supported (books, blogs, trainings) [2].

Table 1.3 Comparative study of the different agile frameworks

Characteristic	XP	Scrum	DSDM	FDD	Crystal
Development	Iterative increments	Iterative increments	Iterative	Iterative	Incremental
Recommended iteration time period	One to six weeks	Two to four weeks	80 % solution in 20% total time	Two days to two weeks	Depending on method from the family
Project team	Smaller teams Less than twenty members	All sizes	All sizes Independent	Many members More than one team	All sizes. Depending on method from the family
Team communication	Informal Daily stand-up meetings	Informal Daily stand-up meetings	Documentation	Documentation based	Informal; Face-to-face
Project size	Smaller projects	All types of projects	All types of projects	More complex projects	All types of projects Depending method from the family
Customer involvement	Customer involved	Costumers through the role of Product Owner	Costumers through frequent releases	Costumers through reports	Costumers through incremental releases
Project documentation	Only basic documentation	Only basic documentation	Documentation exists	Documentation is important	Only basic documentation
Specialties	TDD, user stories, refactoring	Sprint, Product and Sprint backlog, Planning Poker, Scrum master	Prototyping	UML diagrams	Adaptable methods family, all types of projects and team sizes
Advantages	Open workspace, customer as a part of the team, well defined best practices, feedback; Simple written code; Visibility of the whole process; Constant testing; Promotes a highly energizing way of working	High level of communication and collaboration; Team motivation; Transparency (all the team members follow the project process); Focus on quality; Good sprint planning is prioritized, so that the whole scrum team understands the "why, what and how" of allocated tasks	Requirements priority approach, efficient project management	Reports and documentation enable multitasking	Methodologies that adjust to project type and size; Requires a technical environment with automated tests, configuration management and frequent integration; It facilitates closer communication within teams and promotes interaction and knowledge sharing between team members; It requires frequent
Disadvantages	Focus on code can lead to paying less importance to design; This framework may not work at its best if not all the team members are in the same geographical area; Weak documentation; lack of discipline, customer presence is mandatory;	Weak documentation; Not every developer role may be well defined; The segmentation of the project and the search for the agility of development can sometimes lead the team to lose track of the project as a whole, focusing on a single part	Complex documentation	Individual code ownership, not applicable to small projects	Efficient coordination of bigger teams; It might not work best for geographically scattered teams, because of the constant need to communicate and reflect; Planning and development are not dependent on requirements; It is ideal for experienced, autonomous teams.

1.4.3. Scrum framework

1.4.3.1. Definition

The Scrum Framework is a light/simple framework within which individuals, teams, and organizations can tackle complex, adaptive problems while productively and creatively delivering products of the highest possible value [3].

Various processes, techniques, and methods can be employed within the framework. Scrum wraps around existing practices or renders them unnecessary. Scrum makes visible the relative efficacy of current management, environment, and work techniques so that improvements can be made. [3].

1.4.3.2. Scrum theory

Scrum is founded on empiricism and lean thinking. Empiricism asserts that knowledge comes from experience and making decisions based on what is observed. Lean thinking reduces waste and focuses on the essentials.

Scrum combines four formal events for inspection and adaptation within a containing event, the Sprint. These events work because they implement the empirical Scrum pillars of transparency, inspection, and adaptation [**3**].

- ***Transparency***: the emergent process and work must be visible to those performing the work as well as those receiving the work. This transparency creates an open work culture and that is shown through artifacts (project vision statement, prioritized product backlog, release-planning schedule), meetings (sprint review meetings, daily standup meetings) and information radiators (Burn down chart, Scrum board).
- ***Inspection***: The Scrum artifacts and the progress toward agreed goals must be inspected frequently and diligently to detect potentially undesirable variances or problems. To help with inspection, Scrum provides cadence in the form of its five events (*i.e.,* the Sprint, Sprint Planning, Daily Scrum, Sprint Review, and Sprint Retrospective)
- ***Adaption***: If any aspects of a process deviate outside acceptable limits or if the resulting product is unacceptable, the process being applied or the materials being produced must

be adjusted. The adjustment must be done as soon as possible to minimize further deviation.

1.4.3.3. Scrum team

The fundamental unit of Scrum is a small team of people, a Scrum Team (typically 10 or fewer people). The Scrum Team consists of one Scrum Master, one Product Owner, and Developers (Cf. Table 1.4).

Table 1.4 Scrum Team

Role	Description
Product Owner	He is the only responsible on building and managing the product backlog. He ensures that the product is transparent, clear and visible to everyone including the business and the team; and give the development team a clear guidance on what to work on next.
Scrum Master	He is accountable for the Scrum Team's effectiveness. He does this by enabling the Scrum Team to improve its practices, within the Scrum framework. The Scrum Master serves the Scrum Team in several ways, including: (i) Coaching the team members in self-management and cross-functionality; (ii) Helping the Scrum Team focus on creating high-value Increments that meet the Definition of Done; (iii) Causing the removal of impediments to the Scrum Team's progress; and (iv) Ensuring that all Scrum events take place and are positive, productive, and kept within the timebox.
Development Team	They are the people in the Scrum Team that are committed to creating any aspect of a usable Increment each Sprint. The Developers should be self-organized, cross-functional and should have the one-team mentality. They are always accountable for: (i) Creating a plan for the Sprint, the Sprint Backlog; (ii) Adapting their plan each day toward the Sprint Goal; (iii) Holding each other accountable as professionals.

1.4.3.4. Scrum artifacts

Scrum's artifacts represent work or value (Cf Table 1.5). They are designed to maximize the transparency of key information.

Table 1.5 Scrum artifacts

Artifact	Description
Product Backlog	An ordered list of everything needed in a product based on the product goal. This list is always evolving and never complete
Sprint Backlog	A list of everything that the team commits to achieve in a given sprint. Only the development team is responsible for adding to the list.
Increment	Also known as product increment, it is created at the end of every sprint. The team delivers a product increment that meets the agreed-upon definition of done

1.4.3.5. Scrum events

The Sprint is a container for all other events. Each event in Scrum is a formal opportunity to inspect and adapt Scrum artifacts. These events are specifically designed to enable the transparency required (Cf. Table 1.6).

Table 1.6 Scrum Events

Event	Description
The Sprint	It is the heart of Scrum, that has a duration of one month or less, during which a functional and potentially releasable "Finished" Product Increment is created. A sprint starts just after giving the conclusion of the previous sprint. It contains all the work and all the other events that happen during the time-boxed period of development. During a sprint, the objective of the sprint is fixed and the quality objectives maintaining and are never lowered.
Sprint planning	It is an event where all scrum team members meet for maximum eight hours, in order to get the sprint goal and sprint backlog that everyone agrees is realistic and achievable. The product owner can help clarify the chosen product backlog elements and make trade-offs.
Daily Scrum	It is time-boxed to 15 minutes meeting every day. The Daily Scrum is an opportunity for the development team to check-in, assess progress towards achieving the Sprint Goal, and review and plan their activities for the next

	24 hours. Each member of the development team must answer these three questions: – What did they achieve the day before? – What are they going to accomplish today? – What are the obstacles that are holding them back?
Sprint review	The purpose of the Sprint Review is to inspect the outcome of the Sprint and determine future adaptations. The Scrum Team presents the results of their work to key stakeholders and progress toward the Product Goal to be discussed. The sprint review is timeboxed to a maximum of four hours for a one-month Sprint.
Sprint Retrospective	It is time-boxed to a maximum of three hours meeting for a one-month Sprint. The purpose of the Sprint Retrospective is to plan ways to increase quality and effectiveness. This is when the Scrum team reviews what could be improved for future Sprints and how they should do it.

The figure below shows the process of Scrum events:

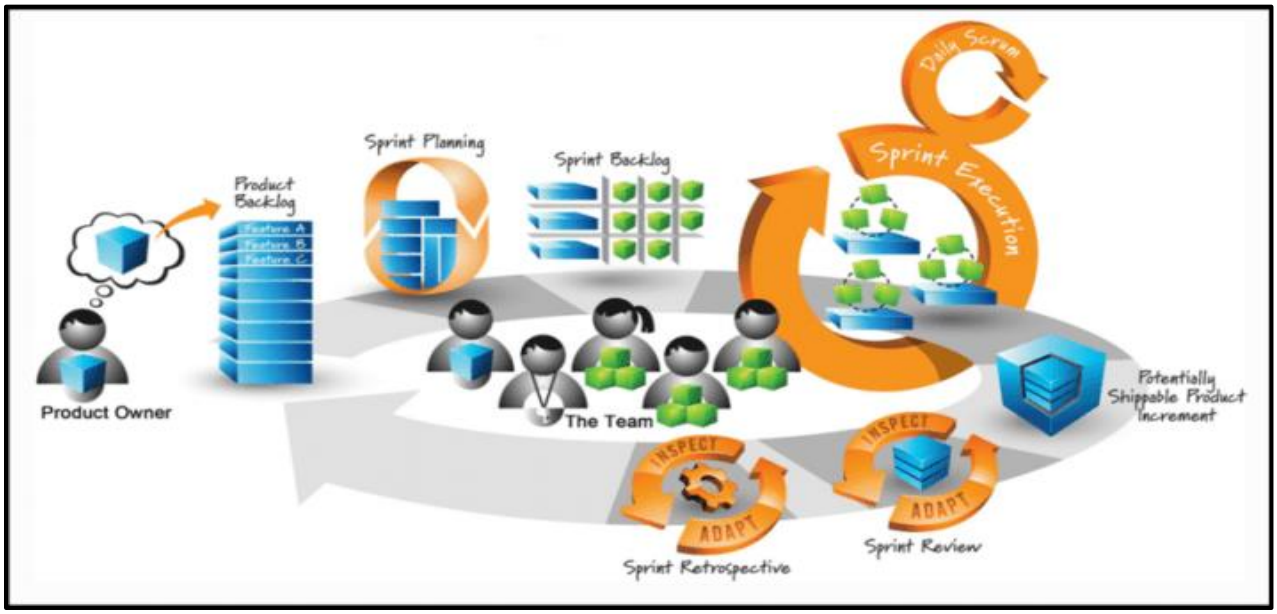


Figure 1.3 The Scrum process: an adaptive, iterative, and incremental cycle

1.5. Conclusion

We have tried in this chapter to clarify and put the project in its general context. We focused in this chapter on the presentation of the host company followed by a study of the existing system in order to specify the objectives to reach, the problematic, and to suggest solutions. Moreover, we have defined the project development and management methodology and the conceptual modeling language that we have adopted.

In the next chapter, we will present the main concepts relating to our field of study.

CHAPTER 2
STATE OF THE ART AND PRELIMINARY STUDY

Chapter 2
State of the art and preliminary study

2.1. Introduction

This chapter presents the basic concepts of our field of study. We first present the domain of Business Intelligence by addressing the concept of data warehouse, ETL, etc. Then, we briefly present the social media that represent our data source. Finally, we define the notion of user profile.

2.2. Business intelligence

2.2.1. Definition

Business intelligence refers to the means, tools and methods that enable the collection, consolidation, modeling and restitution of immaterial data of a company in order to provide decision support and to allow those responsible for the company's strategy to have an overall view of the activity being processed. [4]

2.2.2. BI benefits

Nowadays, several organizations use BI solutions to profit from its great advantages, such as: [5]

- It provides a fast and accurate reporting, analysis, and planning.
- It offers a better data quality.
- It has a wide view of the company statistics
- Business intelligence offers better business decisions.
- It improves employee and customer satisfaction.
- It reduces costs and increases revenues.

2.2.3. Decisional systems vs operational systems

Table 2.1 presents the difference between the operational systems and the decisional systems.

Table 2.1 Comparison between decisional and operational systems.

	Decisional systems	**Operational systems**
Data to manage	Large volumes	Small volumes
Number of users	Limited	Used by the whole company
Process	Open	Closed
Data	Read-only	Read and write
Response time	Medium	So fast
Granularity level	Very large	Low
Database	Centralized	Decentralized

2.2.4. BI architecture

For the success of a BI project, it is necessary to set up different steps in a decision-making chain in order to fully and efficiently benefit from a business intelligence platform. The BI process includes different phases:

- **Data collection phase**: this is the very first phase of the BI process, includes the data collection from the different operational databases, metadata repositories, and external databases (like social media).
- **ETL and data storage**: this phase is about the data cleaning and transforming, followed by data loading in a suitable form for data analysis. It contains in particular the DWH in charge of centralizing the data. It also involves the notions of cubes and data marts.
- **Data visualization**: in this final phase, the usable information is used to create reports, dashboards, etc. It is at this stage that the end-users intervene and analyze the information provided to them. It can also involve specialists in analysis to use statistical tools and come up with forecasts or future estimates (data mining).

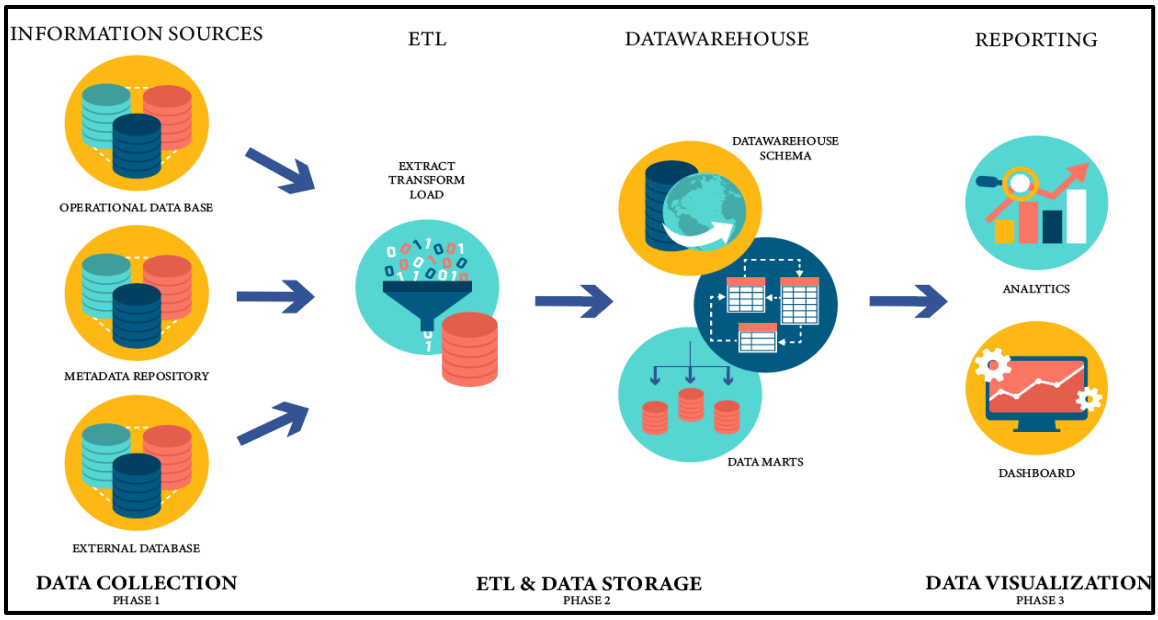


Figure 2.1 BI architecture.

In the following sub-sections, we will detail the most important steps and components that characterize BI.

2.2.5. Data sources

2.2.5.1. Definition

In order to serve the warehouses, the information must be identified and retrieved from their original locations. These are heterogeneous data sources which can be internal company data, stored in the production databases of various departments. They can also be external sources, retrieved via remote services and web services, in addition to sources that can be in flat file format.

2.2.5.2. Social media

The social network is a notion of sociology in which individuals and groups that are interdependent, interact with each other through paths of friendship, family, etc. Thus, relationships and information spread among the members of the network. [6]

Similarly, [7] defines social networking sites as web-based services that allow individuals to build a public or semi-public profile, formulate a list containing users with whom they share a connection, and view and browse this list of connections and those made by other users within the system.

The most popular social media platforms are Facebook, YouTube, LinkedIn, Twitter, Snapchat, Instagram, and more.

- **Facebook** [8] is a sign-up-free website, founded in 2004 by Mark Zuckerberg that allows users to create free profiles, connect with people around the world, and share pictures and videos as well as their own thoughts and opinions.

- **Twitter** [9] is a sign-up-free social networking service, founded in 2006 that allows users post and interact with messages known as tweets.

Figure 2.3 Facebook Logo

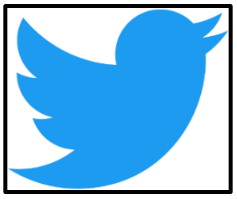


Figure 2.2 Twitter logo

2.2.6. ETL

The ETL processes (Extract, Transform and Load) are in charge of recovering data from all existing heterogeneous operational sources and loading them into the BI system. It is in fact an application able to retrieve the data we need, in their original formats and then integrate them into the global system of our DWH. ETL processes include the following actions (Cf. Figure 2.2):

- **Extract** data from operational databases (ERP, RDBMS, hard files, etc.)
- **Transform** this data by cleaning, conforming, filtering, correcting, joining, splitting, sorting and duplicating.
- **Load** data into the BI system: DWH, DataMart or cube.

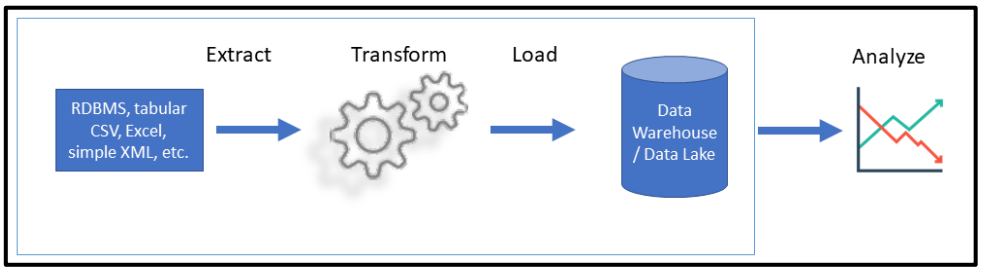


Figure 2.4 ETL process.

2.2.7. Principles of DWH

2.2.7.1. Definition and characteristics

Data Warehouses are databases that store non-volatile structured historical data to enable and support business intelligence activities, perform queries and analysis. [10]

Data warehouses are:

- **Subject-oriented:** data defined and grouped by themes.
- **Integrated**: the data warehouse must be able to collect data from different sources in a consistent format.
- **Non-volatile**: Once a data item reached into the Warehouse, it must not change. We can access the data with read-only mode.
- **Time-variant**: Time variant keys (e.g., for the date, month, time) are typically present.

2.2.7.2. DWH Conception : Data Dimensional Modeling (DDM)

The DWH [11] is composed mainly of facts and dimensions that are linked together according to a well-studied architecture. Dimensional Models have a specific structure and organize data to produce reports that improve performance.

Components of DDM

- **Fact tables:** the fact tables are related to dimension tables with the keys known as foreign keys.
- **Dimension tables:** dimension tables store the Dimensions from the business and establish the context for the Facts. They contain descriptive data that is linked to the Fact Table.
- **Attributes:** those are the various characteristics of the dimension.

Modeling standards

The data warehouse can be modeled in three forms: star model, snowflake model, and fact Constellation model. Each model has its own characteristics. That is why the prediction of the model to be implemented, as a first step, is important.

- **Star model**: This model consists of a fact table in the center, surrounded by dimension tables. The dimension tables are not related to each other (Cf. Figure 2.5).

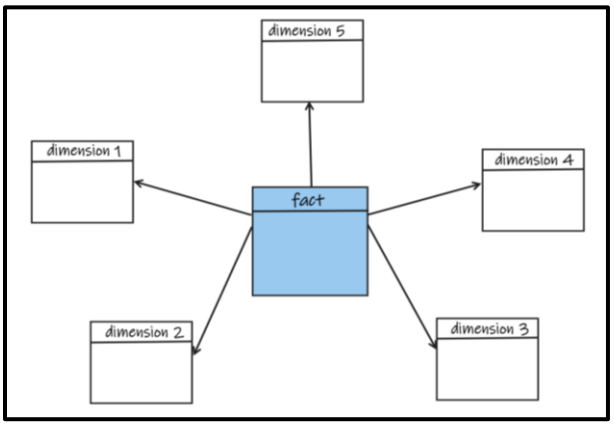


Figure 2.5 DWH - The star model

- **Snowflake model**: This model is a derivative of the star schema where the dimension tables are normalized (the fact table unchanged). The normalization divides the data into additional tables. The principle is to create hierarchies of dimensions (Cf. Figure 2.6).

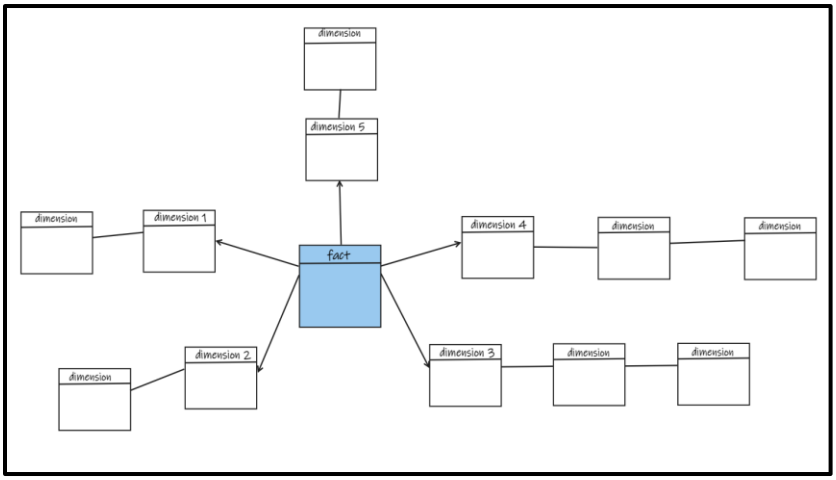


Figure 2.6 DWH - The snowflake model.

- **Fact constellation model**: This model is the fusion of several star models that use common dimensions. It is considered a collection of stars. Multiple fact tables share dimension tables (Cf. Figure 2.7).

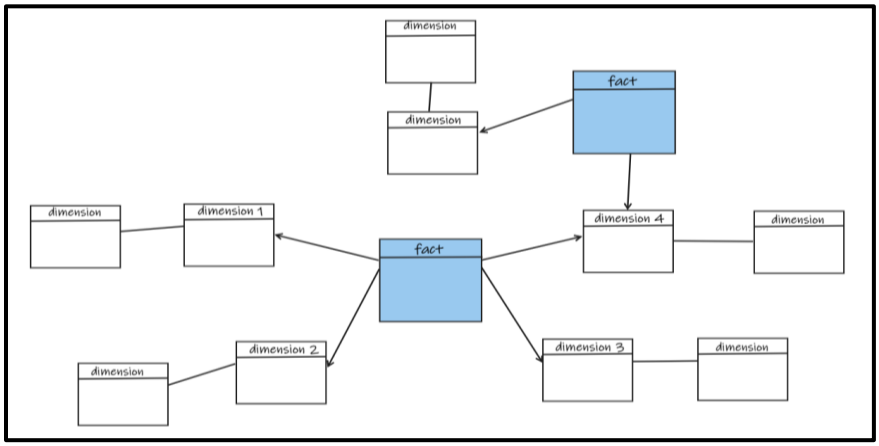


Figure 2.7 DWH - The fact constellation model.

2.2.8. Data marts

Data marts [12] are the small stores that together form the DWH (Cf. Figure 2.8). The data marts can be considered as a sub-set of the DHW. They designed to meet the needs of a particular business sector or function. They represent specific point of view according to business criteria. Data marts are easier to understand and manipulate than DWH, which improve response times.

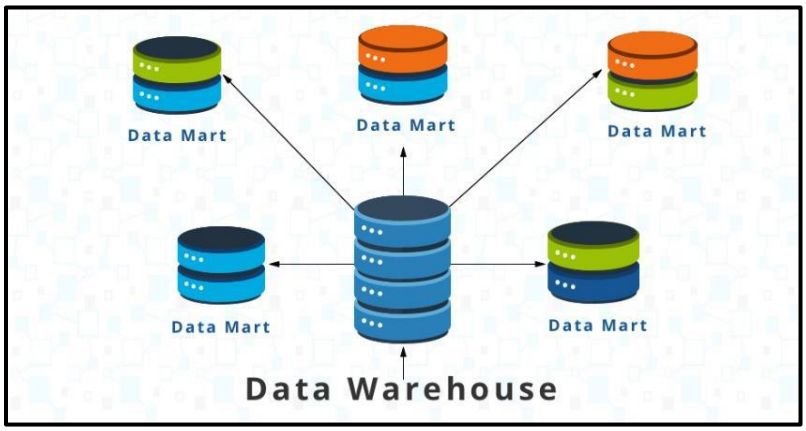


Figure 2.8 Data marts.

2.2.9. OLAP

2.2.9.1. Definition

OLAP [13], stands for Online Analytical Processing, is a computer process that allows users to perform multi-axis data analysis. OLAP provides managers and analysts with fast, interactive access to a set of multidimensional information.

The OLAP system transforms data from the data warehouse or data mart into an OLAP cube. An OLAP cube is a multidimensional database used for fast analysis of data warehouse data. The following figure illustrates an OLAP cube example.

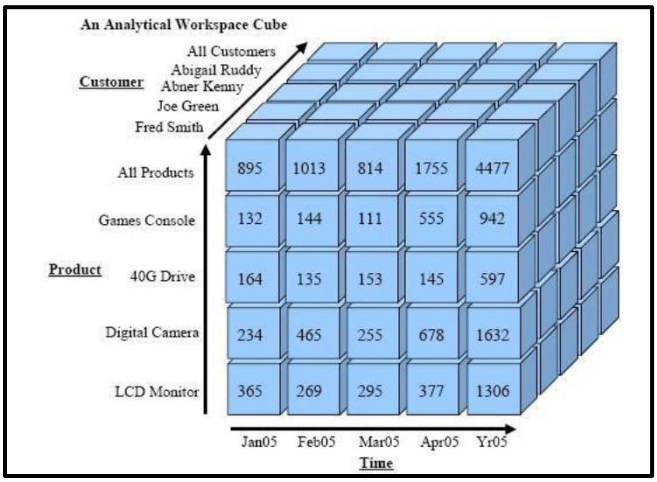


Figure 2.9 Cube OLAP example.

2.2.9.2. Principles of OLAP

The OLAP hypercube provides access to functions for extracting information (for visualization, analysis or processing), and to query functions in MDX language (comparable to SQL for a relational database). These functions are:

- **Rotate**: selection of the target couple of dimensions.
- **Slicing**: extraction of a slice of information.
- **Scoping**: extraction of a block of data.
- **Drill-up**: synthesis of information according to a dimension.
- **Drill-down**: it is the equivalent of a "zoom", the opposite operation of drill-up.
- **Drill-through**: access the basic details of the information.

2.2.9.3. OLAP variants

There are different variants of OLAP. You have the ability to specify how you want your cube to store the data. Each variant has an impact on the performance of the cube.

- **MOLAP**: Multidimensional OLAP stores both data and aggregates in a multidimensional structure. As a result, queries are very powerful and response times are significantly reduced. The disadvantage is that partition processing is resource consuming.

- **ROLAP**: Relational OLAP stores both the data and the aggregates in the source relational database. It is very slow to respond to queries because a ROLAP report is a SQL query. Queries are converted back to SQL. The processing is light because fewer resources are consumed.
- **HOLAP**: The hybrid OLAP stores the data in the source relational database and the aggregates in a multidimensional structure. It is a compromise between MOLAP and ROLAP. It combines the advantages of MOLAP and ROLAP. The response times depend on the queries and the data to be retrieved.

2.2.9.4. Difference between OLAP and OLTP

There are many differences between OLAP and OLTP, some of which are listed in Table 2.2:

Table 2.2 Comparison between OLAP and OLTP.

Parameters	OLTP	OLAP
Process	It is an online transactional system. It manages the modification of the database.	OLAP is an online data analysis and retrieval process.
Features	It is characterized by a large number of short online transactions.	It is characterized by a large volume of data.
Functionalities	OLTP is an online database modification system.	OLAP is an online database query management system.
Method	OLTP uses the traditional DBMS.	OLAP uses the data warehouse.
Requests	Insert, update and delete information from the database.	Mainly select operations.
Data integrity	The OLTP database must maintain the data integrity constraint.	The OLAP database is not frequently modified. Therefore, data integrity is not an issue.
Response time	Its response time is in milliseconds.	Response time in seconds and minutes.
Operation	Allow read/write operations.	Only read and write rarely.

2.2.10. DWH design and modeling approaches

In the state of the art, we find three types of approaches for designing a multidimensional scheme .

2.2.10.1. Top-down approach

In this approach [14], the content of the warehouse is based on the needs of the end-user. The objective of this method is to design a multidimensional schema based on the needs of the decision-makers and the constraints to be respected to ensure the proper functioning of the operational system. (Cf. Figure 2.10).

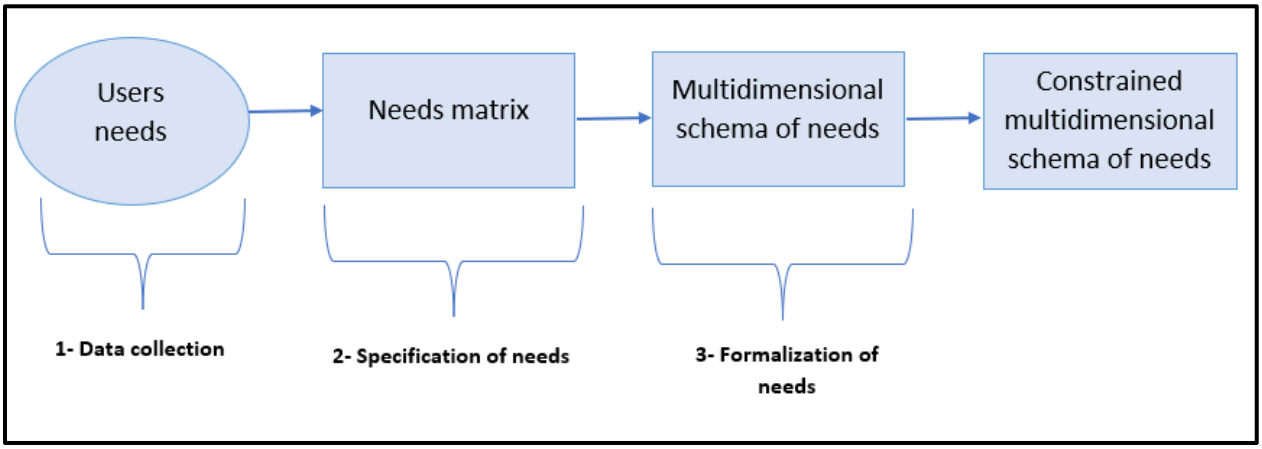


Figure 2.10 Top-down approach

2.2.10.2. Bottom-up approach

This approach is based on the definition of facts and dimensions from the conceptual schema of the data warehouse. This analytical process examines basic data to derive a multidimensional schema that provides an analytical view of the data. (Cf. Figure 2.11).

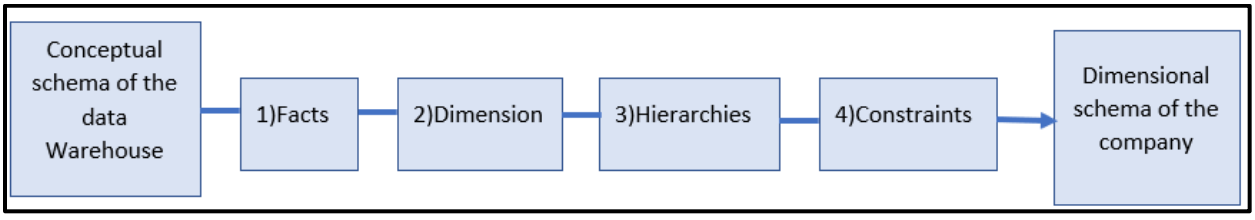


Figure 2.11 Top-down approach

2.2.10.3. Mixed approach

It is a hybrid approach , which combines bottom-up and top-down approaches. In fact, it takes into consideration the data sources and the users' needs.

2.2.11. Data visualization

Data visualization is the practice of converting information into a visual context, such as a map or graph, to make data easier to understand and extract insights from. Data visualization's primary goal is to make it easier to identify patterns, trends, and outliers in large data sets.

2.3. OSINT : Open-Source intelligence

2.3.1. Definition

Open-source intelligence [15] refers to any information about an individual or entity that we can legally collect from free public sources. Practically speaking, this means the information found on the internet, but technically, we can classify any public information as OSINT. This public information can be any book or report, article in a newspaper, or a statement in a press release.

2.3.2. OSINT framework

An OSINT framework helps people find free OSINT resources and tools that focus on gathering information. Figure 2.12 below shows the OSINT framework.

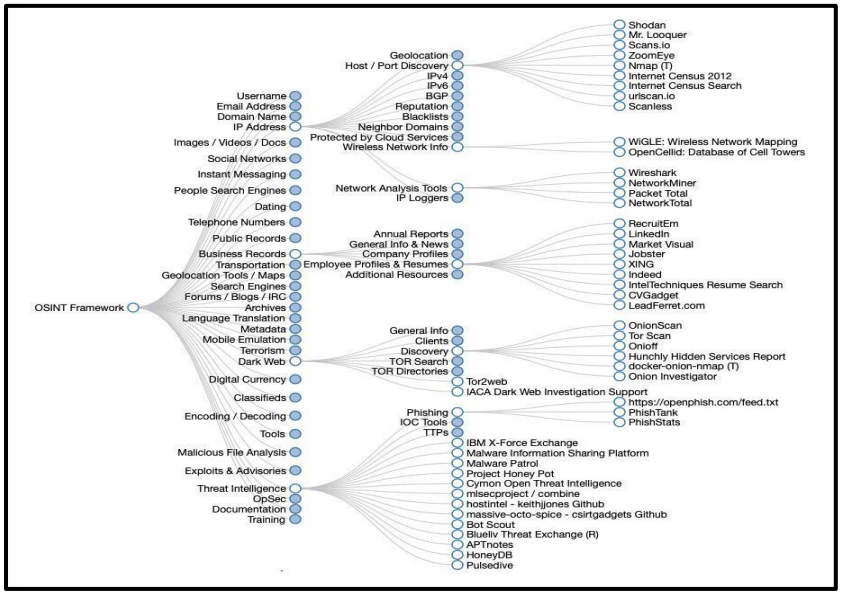


Figure 2.12 Top-down

2.4. User profile

2.4.1. Definition and content

A user profile is a collection of information gathered from most commonly open-source intelligence tools such as social media (Facebook, Twitter, etc.) that is used to characterize the user. The collected information can be demographic data; background, knowledge, skills, goals and needs, behavior, interests and preferences, etc.

A typical user profile may contain information such as a person's name, surname, birthday, gender, geographic location, academic and professional experience, as well as membership in groups. This content can be grouped as:

- *Demographic data*: this category presents very basic information like name, gender, age, native language, country, relationship status, etc.
- *Background knowledge and skills*: this attribute presents the education history of the user, the professional roles the user has expertise within a company, the user skills, etc.
- *Goals and needs*: represent what the user wants to achieve in a given context. We can find this information in the user browsing history, groups and pages he/she follows, information the user has on his profile, etc.
- *Behavior*: not all users are made equal, the behavior of each is represented by the user history of a repetitive action due time of the day or day of the week, or any other information like what kind of device he is using, what device OS, in which geo-location, etc.
- *Interests and preferences*: represent the user's professional interests, hobbies, interests in entertainment, interests in sports, etc.

2.4.2. User profile phases

According to Marina et al. [16], a user profile has three main phases:

- Data collection is the phase where it is all about obtaining and extracting user data. The data can be gathered either by using manual techniques like registration forms, questionnaires, etc. in which we use effort from the user, this method is called explicit; or by using other techniques like gathering information from user search logs, browsing history, etc. that don't use any effort from the user, and this is the implicit method. Moreover, the hybrid method, combine both implicit and explicit sources.

- Building and modeling user profile is the phase that is concerned with the storage and the representation of the information collected in the first phase. There exist three types of user models which are vector-based, semantic networks, and concept-based. Vector-based user model or also called a vector-space model is a collection of data referred to as a set of terms by weighted vectors of keywords. Semantic network-based is about mapping the words of concepts and interests. It presents the profiles by the weighted concept of each node in the semantic network. Moreover, nodes of concepts and relationships represent concept-based profiles.
- User profile personalization is the phase that assumes exploiting information in personalized services such as recommender systems, social networks, and browsing.

The following figure presents a classification of the phases of user profiling.

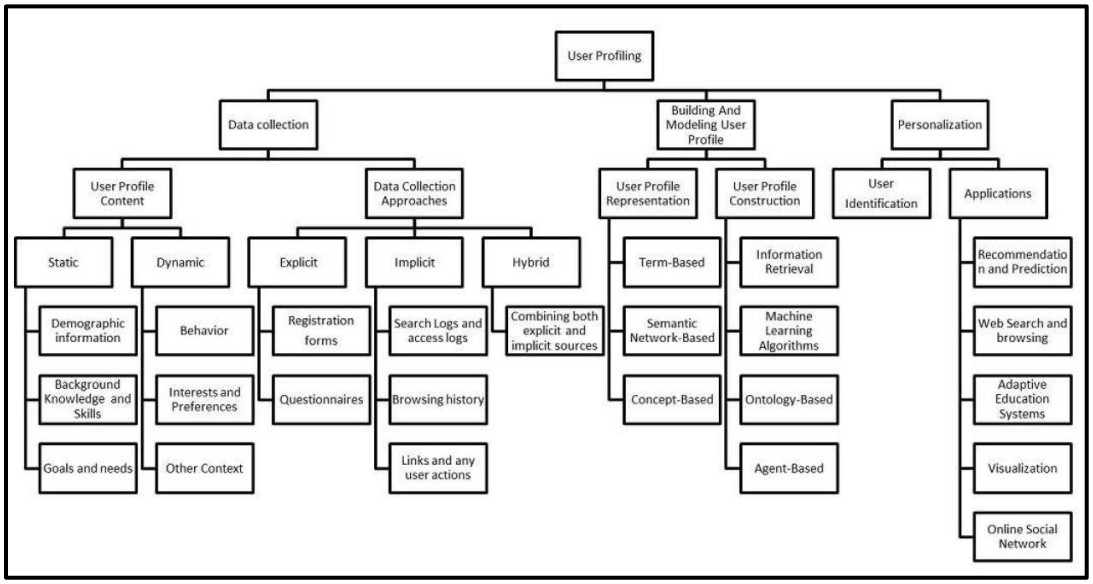


Figure 2.13 User profiling phases

In our case, we will be collecting public and available user demographic data and user behavior, using the implicit method, through collecting the user interaction with posts in groups he/she follows. Furthermore, we will build the user profile using machine-learning algorithms to finally visualize and recommend the best profiles that matches with a certain product produced by the company.

2.5. Conclusion

Through this chapter, we have detailed all the notions related to the decisional systems following by the study of each phase of the business intelligence phase, as well as the definition of a user profile. Indeed, these notions will be the basis for the realization of our solution within the frame of our project.

CHAPTER 3

SPRINT 0: SPECIFICATION OF NEEDS AND ELABORATION OF THE PRODUCT BACKLOG

Chapter 3
Sprint 0: Specification of needs and elaboration of the product backlog

3.1. Introduction

As previously stated (in the first chapter), we will manage our project using the Agile-Scrum framework. The Sprint 0 is the most important phase in the Scrum development cycle because it directly influences the success of the sprints, particularly the first. The work done during this time period leads to the definition of the Scrum team, the creation of a good product vision, the identification of user stories for the product backlog, and the planning of sprints.

Sprint 0 is the time spent preparing for the completion of a project. Unlike the other sprints in the project, Sprint 0 has no set duration. Sprint 0 is a project-oriented sprint. In addition to developing the product backlog, this sprint aims to identify the technological and architectural specifications of our project.

3.2. Scrum team

The SCRUM roles for our project were as follows (Cf. Table 3.1 and Figure 3.1):

Table 3.1 Scrum team

Product Owner: Mrs. Rodile ANNABI	This is the product's owner. He is a general expert in the field.
Scrum Master: Mrs. Amel BOURASSI	This is the team leader: The person who ensured that my project's various sprints went perfectly.
Scrum Team: Abir ALOULOU	It is the development team. It is in charge of the application's analysis, design, and technical organization.

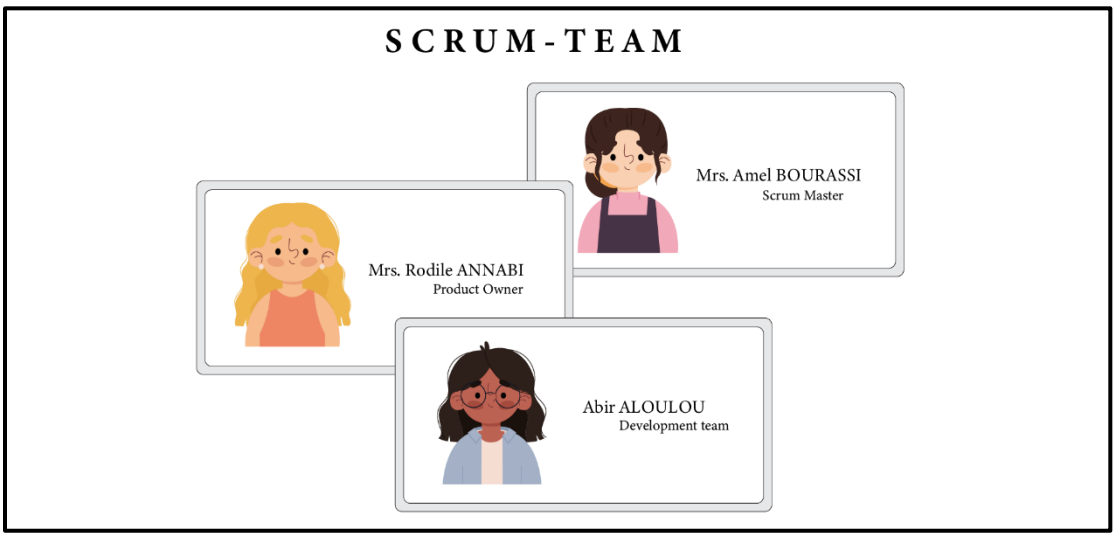


Figure 3.1 Scrum team.

This phase allowed us to integrate into a team that served as a space for reflection and discussion of project information. This space may require environment adaptation to provide a pleasant and communication-oriented work environment within the team.

3.3. Conceptual modeling language: UML

UML [17], Unified Modeling Language, is a notation allowing to model a system in a standard way. This language was created by combining several previously existing methods, and it quickly became the standard for object modeling.

UML can be associated with any computer system design process, at any stage of the process, and with different programming environments.

In fact, UML is not a method, but a language with a well-defined syntax and rules that try to achieve the described goals through a formed graphical representation.

UML [17] represents then communication support between customers and designers, as well as between teams of designers. For the design of our application, we exploit several UML diagrams such as use case diagrams, class diagrams, and sequence diagrams.

3.4. Product Backlog elaboration

The most important Scrum artifact is the product backlog, which is the set of functional or technical characteristics that comprise the desired product. The functional characteristics are referred to as user stories, while the technical characteristics are referred to as technical stories.

The steps presented here for developing the product backlog will assist in answering the following three questions:

- What features will be created?
- Should they be delivered in what order?
- To whom are they addressed?

3.4.1. Defining the project's vision

The objective of this step is to define the objective that our project "Analysis of user profiling and behavioral data in the medical sector" must achieve before the product owner stops development.

Thus, we aim at proposing a business intelligence solution for profiling members of various Facebook groups dealing with the same subject based on their connection data and their behavior on the web.

The functional requirements for this project have been classified into these themes:

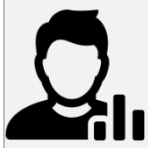
- Data collection from different public sources,
- Processing: Cleaning the data and converting it to a standard format,
- Data storage in a centralized location,
- Exploitation of the collected database: data point collection analysis,
- Obtaining additional data,
- Produce a report.

3.4.2. Actors identification

An actor is a physical entity, a person, or an abstract entity, software, capable of interacting with the system in order to satisfy a defined need. Therefore, it is essential to identify the actors, so we will delimit and identify who is dedicated to this project.

We identify three actors in our system, as illustrated in table 3.2:

Table 3.2 Actors' identification

Icon	Actor role	More information
	Data analyst	• **Post definition:** A person capable of doing reports by turning information into insights, so he/she can use it in making business decisions and predicting future trends. • **Description:** The data analyst uses the system to make reports from useful information according to the company needs, and to conclude business decisions from them. • **Satisfaction requirements:** The data analyst wants to receive a real-time data, and wants the system to be fast and easy to use. • **Business value:** High • **Use frequency:** Regular • **Technology knowledge level:** Medium
	Marketing manager	• **Post definition:** A person responsible for developing, implementing, and executing strategic marketing plans for an entire organization to attract potential customers and retain existing ones • **Description:** According to the reports done by the data analyst, the marketing manager use the system to give helpful marketing strategies for the company's products benefit. • **Satisfaction requirements:** The marketing manager wants to receive up-to-date reports and wants the system to be fast and easy to use. • **Business value:** High • **Use frequency:** Regular • **Technology knowledge level:** Medium
		• **Post definition:** A person who supervises and leads a company's operations and employees in order to ensure

	Business manager	company productivity and efficiency including implementing business strategies, evaluating company performances, and more. • **Description:** According to the business decisions and the marketing strategies done by the data analyst and the marketing manager, the business manager use the system to evaluate the company's performance and give business strategies. Furthermore, the business manager uses a system to always keep an eye on the dashboard results, so he can make sure of the decisions made by the team. • **Satisfaction requirements:** The business manager wants to receive up to date reports, and wants the system to be fast and easy to use. • **Business value:** High • **Use frequency:** Regular • **Technology knowledge level:** Medium
	BI modeler	• **Post definition:** A person who is in charge of managing data retrieval and analysis within a company. • **Description:** A BI modeler is the one responsible for extracting the data, transforming it into useful information, and loading it into a centralized database so that the data analyst can use it in future reporting. • **Satisfaction requirements:** The BI modeler wants to have a list of requirements so that he can build the process. • **Business value:** High • **Use frequency:** Regular • **Technology knowledge level:** Medium

3.4.3. Specification of requirements

Our application must satisfy the requirements of all the actors involved. We expose the following functional needs as well as non-functional needs.

3.4.3.1. Functional requirements

The functional requirements section defines what our product must do.

The system must enable:

- **From a data warehouse perspective:**
  - To build a non-volatile data source.
  - To make the database more scalable and more adapted to future needs.
  - To keep the traceability of each data: keep the source of the extracted data.
  - To manage a large volume of varied data.
  - To build an automated system for the data flow.

The application must permit:

- **From the application point of view:**
  - Receive real-time data to be able to make up-to-date reports.
  - Access to the database to be able to perform analysis to assess the quality and meaning of data and prepare reports.
  - Easily access the dashboard and consult the visual reports.
  - Consult the visual reports to be able to make marketing strategies.
  - Receive up-to-date reports to be able to make business decisions.
  - Satisfy the needs and interests of the Facebook group member.

The system should easily generate reports as required:

- The number of profiles per group category.
- The number of likers in a certain week, per group.
- The number of commenters per group.
- The number of employed active users.

3.4.3.2. Non-functional requirements

The non-functional requirements part must describe the general proprieties of the system including how it should be developed and maintained, in order to fulfill all functional needs.

The non-functional requirements that our system must meet are as follows:

- **Security:**
  - The system must be reliable and secure.
  - The system must ensure the security and the performance of the data warehouse.
- **Maintainability and scalability:**
  - The code of our application must be clean, readable, and understandable to allow easy maintenance and improvement in the future.
  - It is possible to add other modules to make the system evolve. It is necessary to take into account the possibility of its extension by adding new functionalities.
- **Ergonomics and usability:**
  - The system must provide clear and easily analyzed dashboards.
  - The system must guarantee an interactive, user-friendly, understandable, and easily manipulated dashboard.
- **Performance and utility:**
  - The system must deliver useful, accurate, and precise results.
  - The system must serve all users without overload. The delivery of massive data must always be functional by ensuring fast processing and quick response time to provide the requested reports.
- **Software documentation:**
  - Documentation can be for professionals or the final users of the software. Indeed, it is necessary to write a text that explains to any user the steps to follow to use the application produced.

3.4.4. Product backlog

Scrum does not require a specific practice for identifying and naming backlog items. The most typical example is to refer to an element as a story or a use case. The user stories in a product backlog are classified in the order in which they will be completed. This concept of priority is extremely crucial in iterative development.

The following table represents our project's product backlog, which includes features, user stories, priorities, rank and effort (velocity).

More precisely, each user story is assigned a rank based on its risks. We begin the treatment of our user stories with the highest priority and lowest risk use cases.

Each user story has an effort (velocity) in addition to the rank, which is an initial estimate of the amount of work required to implement this requirement. This effort is measured in story points, which are equivalent to ideal man-days. In general, the one-story point is equivalent to one man/day.

Table 3.3 Backlog Product.

ID-F	Features	ID-US	User story	Priority	Effort
1	Authentication	1.1	As a BI modeler, data analyst, business manager, and marketing manager I want to be able to log in and access my sections of the application.	Must have	1
2	Create the multi-dimensional mod of the DWH	2.1	As a BI modeler, I want to create the multi-dimensional model of the DWH in order to load/store useful information in a centralized and non-volatile database	Must have	6
3	Feed the DWH	3.1	As a BI modeler, I want to extract data from various sources in order to collect all the public information of users in the target groups.	Must have	3
		3.2	As a BI modeler, I want to transform and clean data in order to convert the data into useful information.	Must have	4
		3.3	As a BI modeler, I want to load/store useful information in a centralized and non-volatile database (i.e., DWH).	Must have	2
		4.1	As a data analyst, I want to have real-time data to make rigorous decisions.	Must have	1

4	Analyze and exploit the data/the relevant information	4.2	As a data analyst, I want to create and configure reports depending on the company's needs.	Must have	1
5	Make business decisions	5.1	As a business manager, I want to have a real-time data to make rigorous decisions.	Must have	1
		5.2	As a business manager, I want to easily access the dashboard and consult the visual reports in order to make business decisions.	Must have	2
6	Make marketing strategies	6.1	As a marketing manager, I want to easily access the dashboard and consult the visual reports in order to make marketing strategies.	Must have	1
7	Receive Facebook Ads	7.1	As a Facebook group user, I want to get useful ads according to my interests.	Should have	1

3.5. General Use Case Diagram

Use case diagrams are UML diagrams used to provide a high-level overview of the features offered by our application to the user (Cf. Figure 3.2).

In our decisional system, we have the following actors: The Facebook group user, the data analyst, the marketing manager, and the business manager.

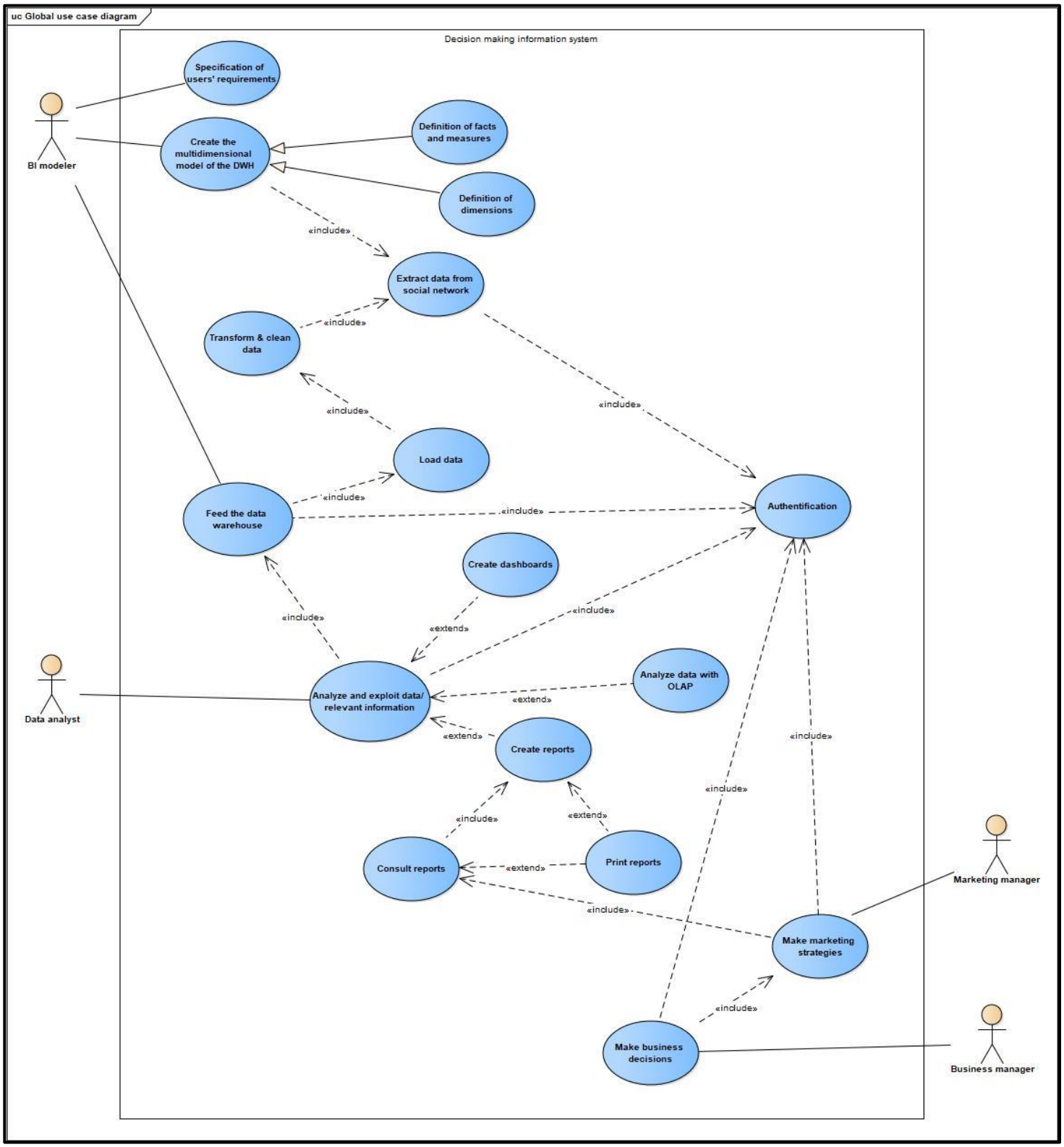


Figure 3.2 General use case diagram.

3.6. General sequence Diagram

The following figure illustrates the sequence of events provided by the system to provide users with reports and dashboards that can assist them in decision-making.

The global sequence diagram describes the main steps of our solution. Indeed, the first step consists in extracting the data from “Facebook” using “Phantom Buster”, which will then be stored in our “PostgreSQL” database as an intermediate step, detailed in Sprint 1. Then, the gathered data will go through the "TL" (Transform Load) processing in which we will use “Talend” and python script for the data cleaning and “SSMS” for data loading. Finally, the data stored in “MSSQL Server” is sent to “Power BI Desktop” for purposes of data visualization.

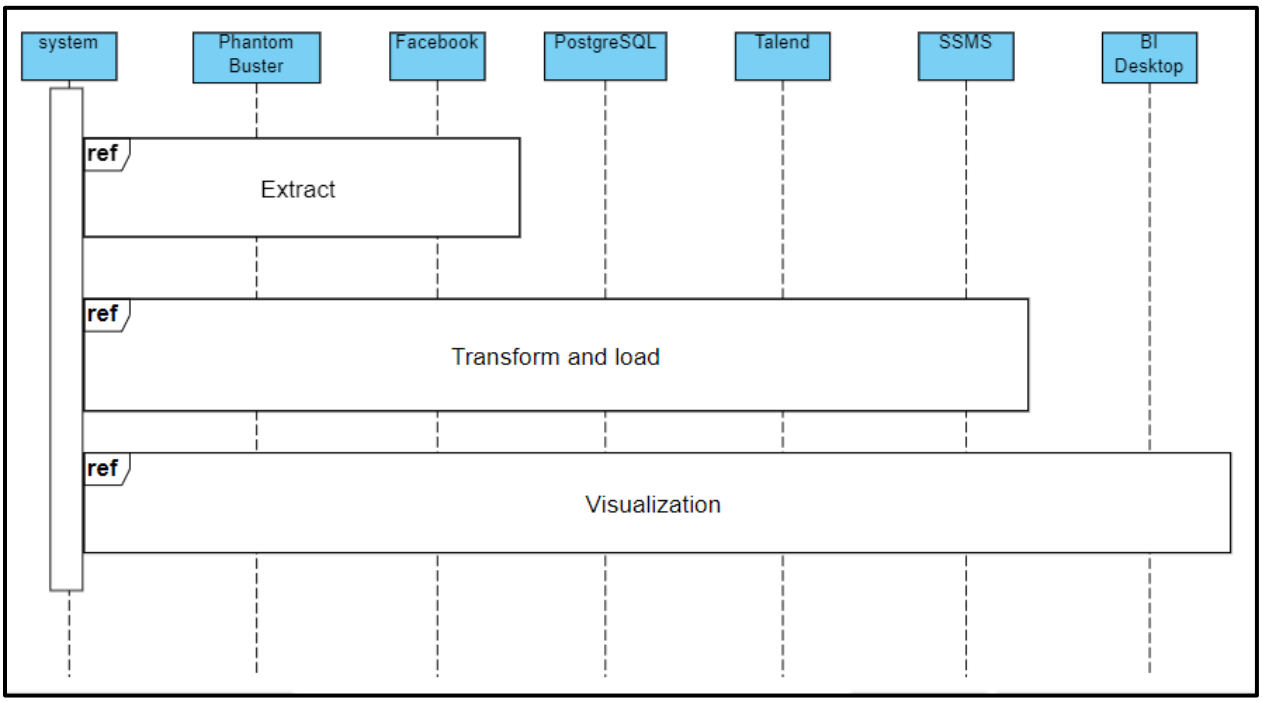


Figure 3.3 General sequence diagram.

3.7. Division of the Project

Our project is structured by dividing it into different batches of activities in order to have sub-parts whose complexity is more easily controllable (Cf. Figure).

The Scrum framework necessitates the division of the system into Releases. A Release is a series of Sprints that culminates in a product that provides sufficient value to its users.

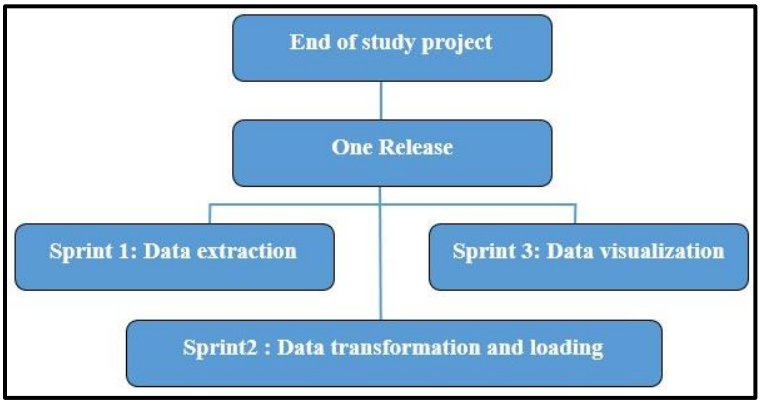


Figure 3.4 Breakdown of the project

3.8. Sprint planning

The sprint planning meeting is the most important event in Scrum. The purpose of this meeting is to plan the work schedule and identify the sprint backlog, as shown in figure 3.5.

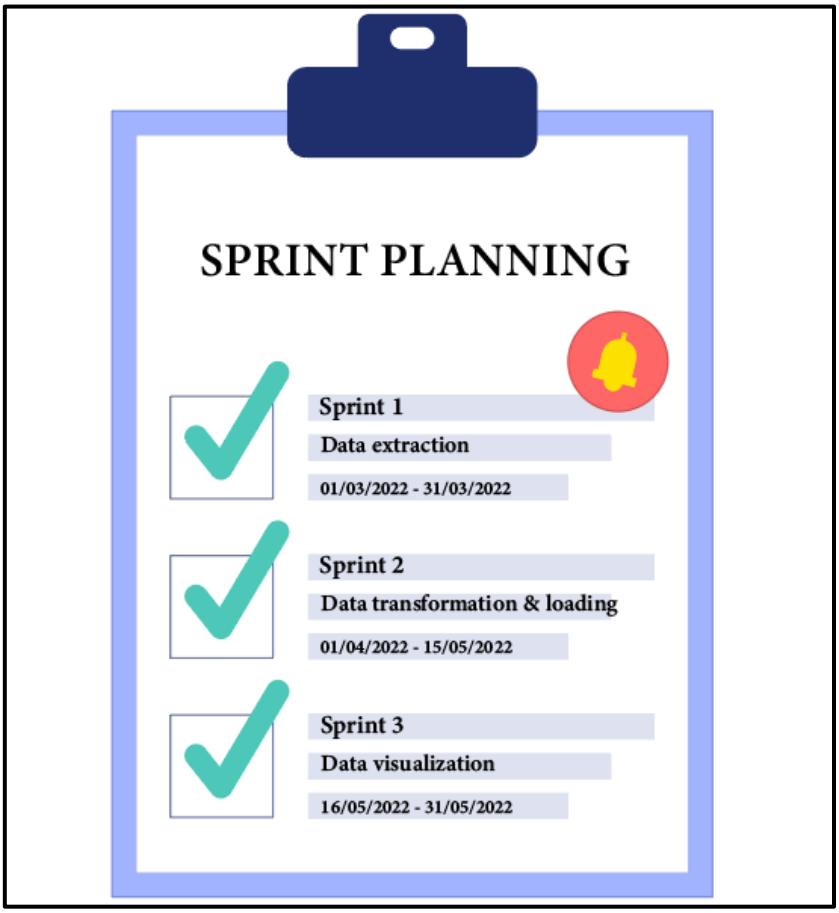


Figure 3.5 Sprints planning

3.9. Specification of the working environment

The development environment refers to the set of tools and languages used in the implementation of the project solution. This part will cover the hardware and the software details.

3.9.1. Hardware environment

In the development of this system, we used a laptop that contains the following characteristics:

Table 3.4 Laptop Characteristics

Brand	**Dell**
Processor	Intel® Core ™ i3-4005U CPU @ 1.70GHz
RAM	8,00 Go
Operating system	Windows 10 professional

3.9.2. Software environment

Choosing the software environment is a very important step in the project development. For this purpose, we have studied, compared and tested several software programs in order to establish a suitable environment for a BI.

3.9.2.1. Phantom Buster

Phantom Buster [18] is a technology company created in 2016. It offers data scraping and automation solutions in form of phantoms for over 20 categories online including Facebook, twitter, LinkedIn, etc. It has a trial version that lasts for 14 days with a limited usage.

Figure 3.6 Phantom Buster logo

3.9.2.2. Visual Code Studio

Visual Studio Code [19], also known as VS code, is a lightweight but powerful source-code editor developed by Microsoft that supports a great number of languages such as Python, Java, and PHP thanks to its available extensions. It is available for Windows, macOS and Linux. People prefer VS code because it integrates with build and scripting tools to perform common tasks making everyday workflows faster.

Figure 3.7 Visual code studio logo

3.9.2.3. PostgreSQL

PostgreSQL [20], also known as postgres, is a relational database management system that is free and open-source, with an emphasis on extensibility and SQL compliance. It is available for macOS, windows, Linux, FreeBSD, and OpenBSD. PostgreSQL helps developers store large and complex data safely, run administrative tasks and create integral environment.

Figure 3.8 PostgreSQL logo

3.9.2.4. Talend

Talend [21] is an open-source data integration platform that was founded in 2005. It offers data integration, data management, enterprise application integration, data quality, cloud storage, and Big Data software and services.

Figure 3.9 Talend logo

3.9.2.5. SSMS

SQL Server Management Studio [22] is a software application that was introduced with Microsoft SQL Server 2005 and is used to configure, manage, and administer all components of Microsoft SQL Server.

Figure 3.10 SSMS logo

3.9.2.6. Power BI Desktop

Power BI Desktop [23] is a free application, developed by Microsoft Corporation, built for the analyst. It gives its users the ability to connect different data sources, combine, transform, visualize the data, and create and share reports.

Figure 3.11 Power BI logo

3.9.2.7. Enterprise Architect

Enterprise software architecture [24] can be used to reduce system complexity and thus improve overall system efficiency. Organizations looking to improve major IT systems look for specialized enterprise IT designs. Enterprise software architects consistently aim to increase system agility by refactoring existing solutions.

Figure 3.12 Enterprise architect logo

3.9.3. Programming language: Python

Python [25] is a high-level, interpreted, object-oriented programming language mainly used in machine learning and data science. It has strongly contributed to the development of big data thanks to its numerous libraries such as Pandas, Numpy, psycopg2, etc.

Figure 3.13 Python logo

3.10. Conclusion

In this chapter, we have described the sprint backlog. In the first section, we elaborated on our project's product backlog and then planned our sprints. Second, we selected the appropriate hardware and software environment.

The following chapter describes the first sprint of our project.

CHAPTER 4
SPRINT 1 : DATA EXTRACTION

Chapter 4
Sprint 1 : Data Extraction

4.1. Introduction

This chapter will concentrate on the development process of the first sprint, with the goal of explaining the first step of data extraction in the ETL process. This sprint includes several functionalities, such as authentication and data extraction.

We first elaborate the sprint backlog. Then we present the requirements analysis phase, which consists of creating the sprint's use case diagram, followed by a textual description of the various functionalities. The sequence diagram representing the interactions between the actors and the BI system is then presented. Finally, we present the realization phase.

4.2. Backlog sprint 1

The Sprint Goal, the Product Backlog items selected for the Sprint, plus the plan for delivering them are together referred to as the Sprint Backlog.

We reviewed the product backlog during the sprint planning meeting with the product owner and scrum master to determine which features are prioritized and which will be integrated into this first sprint. Our objective in this sprint consists in performing the first phase of the ETL process relating to data extraction from Facebook group members using Phantom Buster tool.

Table 4.1 Sprint Backlog -Sprint 1

ID-US	User story name	User story description	Tasks	Week
3.1	BI Modeler-Extract data from various sources	As a BI modeler, I want to extract data from various sources in order to collect all the public information of users in the target groups	Specification of users' requirements	W1/W2
			Identify data sources	
			Determine the structure of the data generated from each source	
			Define the schema describing the link between the data	
			Determine the sequence of extraction tasks	

4.3. Requirements analysis

During this step, we identified the sprint's actor and described the various features/user stories with their use cases.

4.3.1. Use case Diagram

In the following we present the use case diagram "sprint1- data extraction", along with a textual description of each case/functionality.

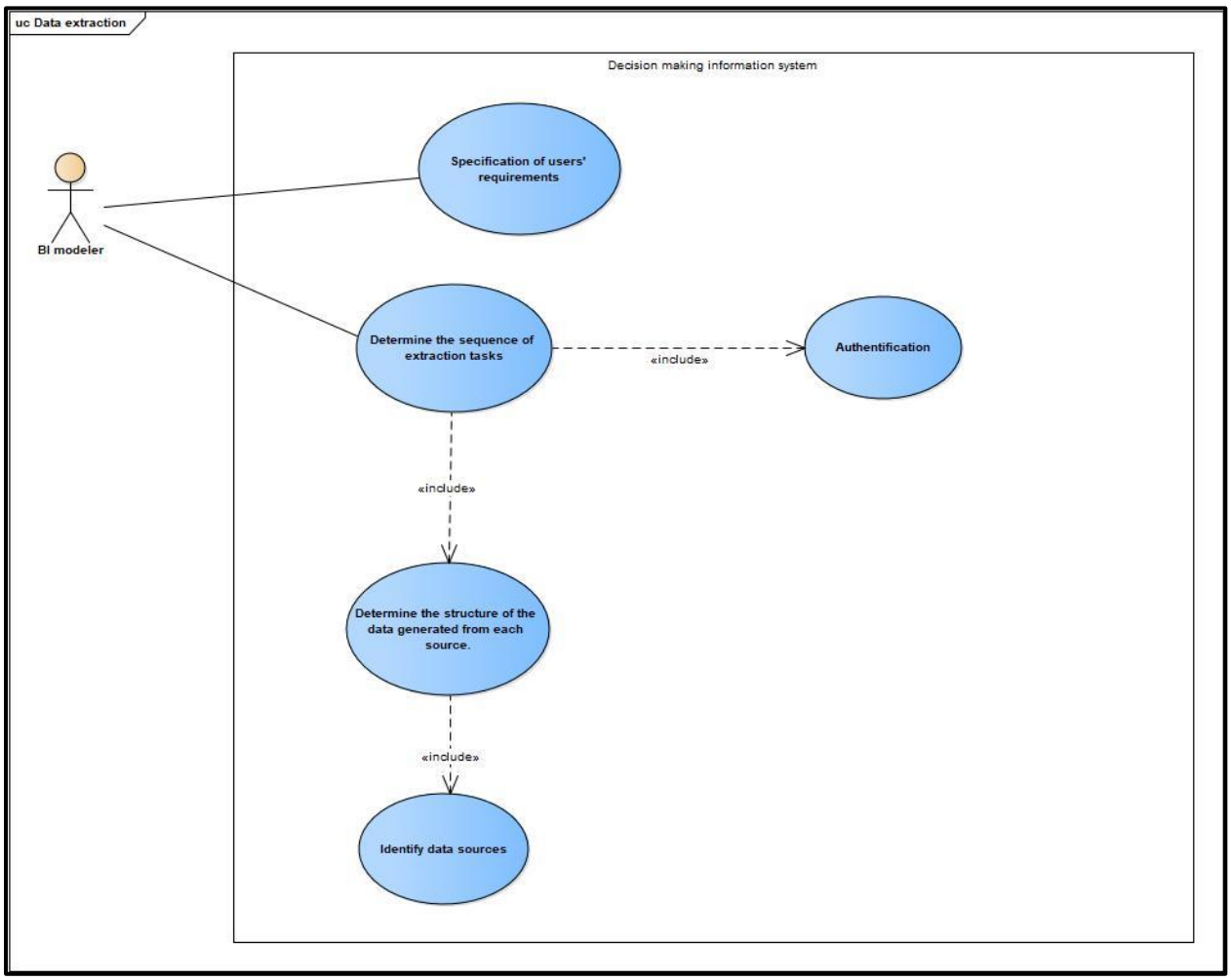


Figure 4.1 Use case diagram-Sprint 1

4.3.2. Textual description of use cases

Table 4.2 Textual description of the use case "Data extraction"

Title	**Data extraction**
Resume	The system allows the BI modeler to extract data from different sources.
Actor	BI modeler
Precondition	The extractor (BI Modeler) has the right to access the data sources
Nominal Scenario	BI Modeler analyzes the customer's needs. BI Modeler determines the data needed for the BI solution. BI Modeler identifies the data sources and their structures. BI Modeler defines the schema describing the link between the data. BI Modeler determines the sequence of extraction tasks.
Post-condition	Extracted data.
Complements	BI Modeler must determine the necessary data to extract.

4.4. Conception

The second activity in a sprint is conceptual modeling, which is translated by the sequence diagram and the class diagram. We chose the sequence diagram to represent the interactions between the actors and the system in chronological order in this work.

4.4.1. Sequence diagram

The data extraction is launched by the system that sends a request to phantom buster to gather the data from Facebook. In its turn, phantom Buster tries to get access to Facebook data using the Facebook API.

If Facebook gives access to Phantom Buster, Phantom Buster will return a message to the system and ask the API Facebook to collect the data. Afterward, Phantom Buster returns the data in a JSON format to the system.

Else, if Facebook does not give access to Phantom Buster, Phantom Buster will return a “Failed access” message to the system.

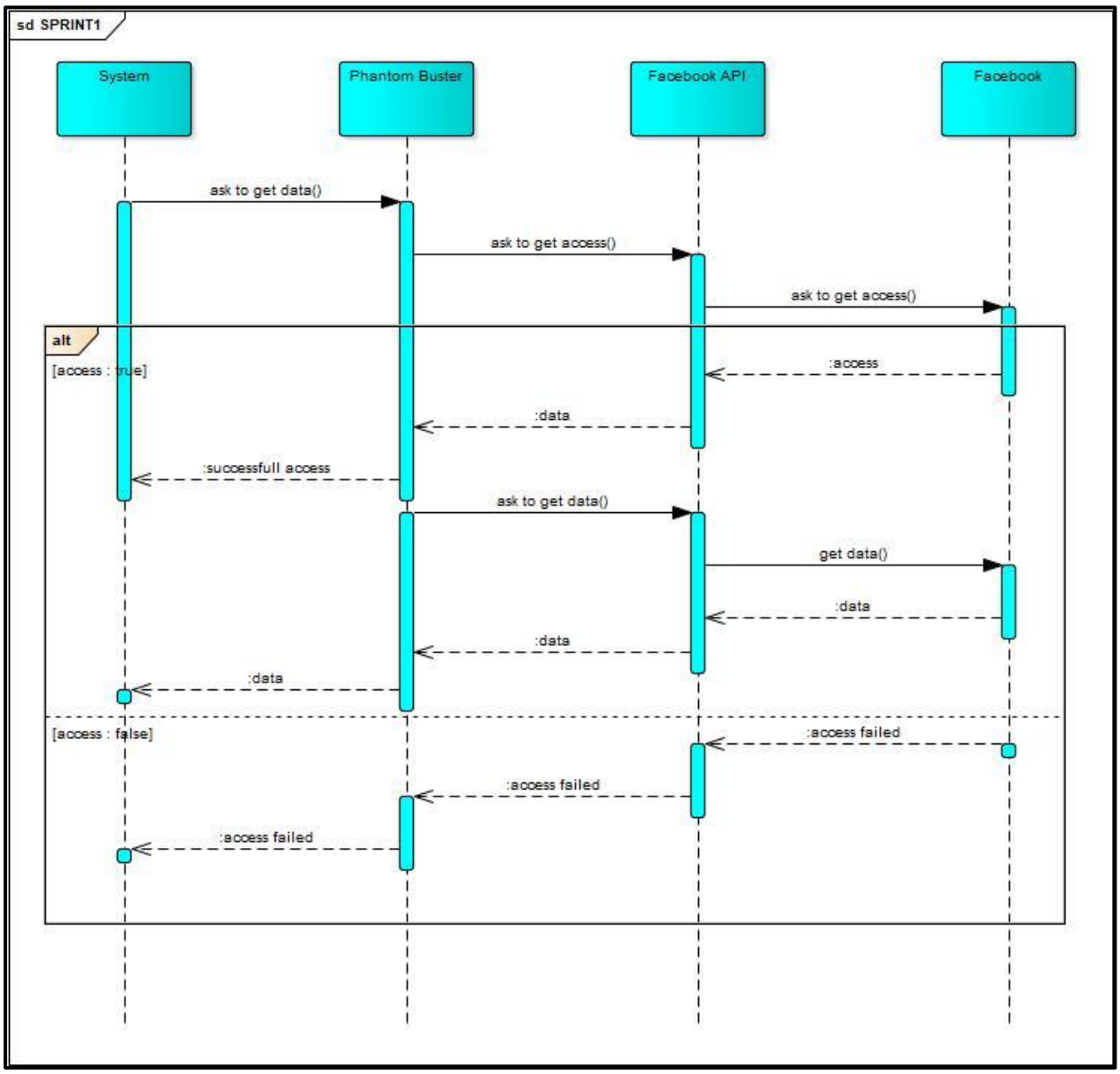


Figure 4.2 Sequence diagram- Sprint 1

4.5. Realization

4.5.1. Data source and extracting tools study

This part might be the most important part in the extraction phase as it contains the data source from which we will be extracting the data and the tool we will be using to extract that data. That is why we need to choose wisely what we are going to work with.

4.4.1.1. Data source

Our project aims on extracting public and open-source data from social media because people nowadays uses social media more than anything and so it contains huge amount of user information.

The table blow presents the huge number of users and data worldwide in different social media platforms, according to statistics made between 2019 and 2022:

Table 4.3 The Quantum of users and data in different social media

Source	Quantum of data
Facebook	Monthly active users of January 2022: 2910 million users Total number of daily users: 1929 million users (the 4th quarter of 2021) Number of user data requests issued to Facebook from U.S. federal agencies and courts: 61.262 (2nd half of 2020)
YouTube	Monthly active users of January 2022: 2562 million users Daily active users worldwide: around 350 and 375 million (4th quarter of 2021)
Twitter	Monthly active users of January 2022: 436 million users Total number of tweets sent per day: 500 million tweets (last updated 21/02/22)
Instagram	Monthly active users of January 2022: 1478 million users Number of daily active Instagram stories users: 500 million (January 2019) Number of sponsored influencer posts: 6.12 million posts (2020)

Some statistics about the social media users, applications, and impact found in Statista [26] [27] [28] and statcounter [29] are presented in the following figures.

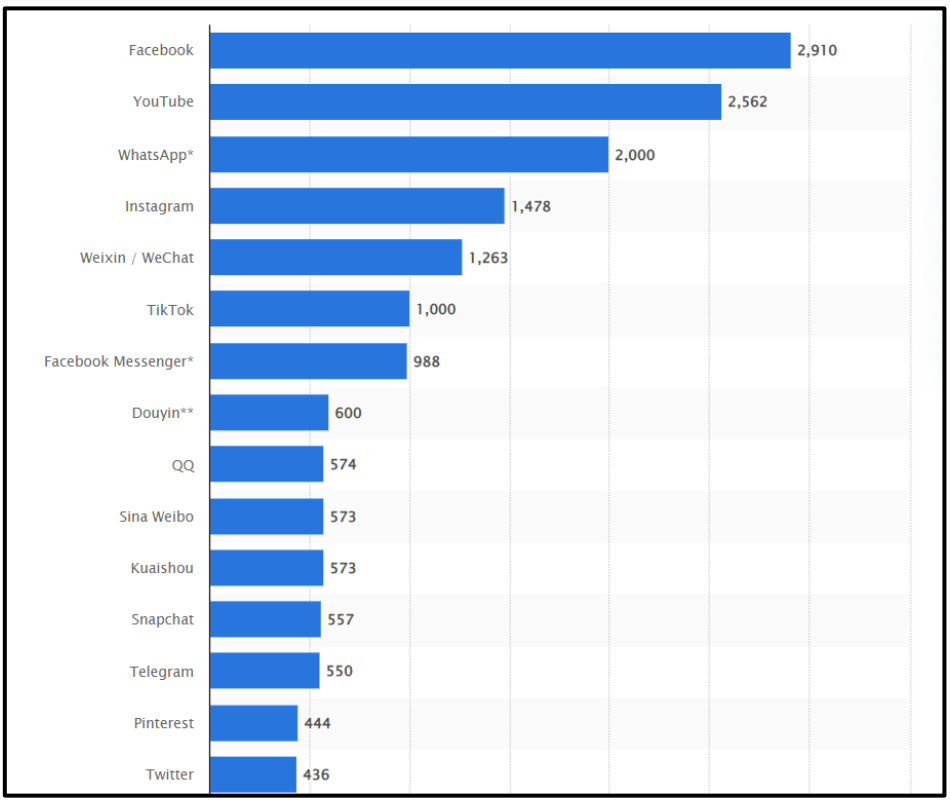


Figure 4.3 Statistics for the number of users of the most popular social media worldwide in January 2022

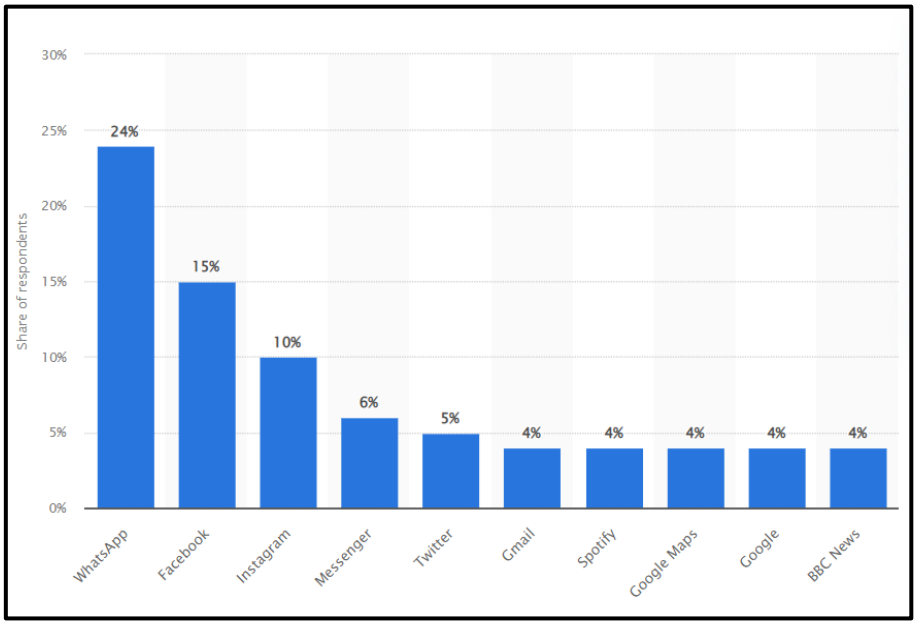


Figure 4.4 Statistics of the most important mobile apps in the UK in 2020

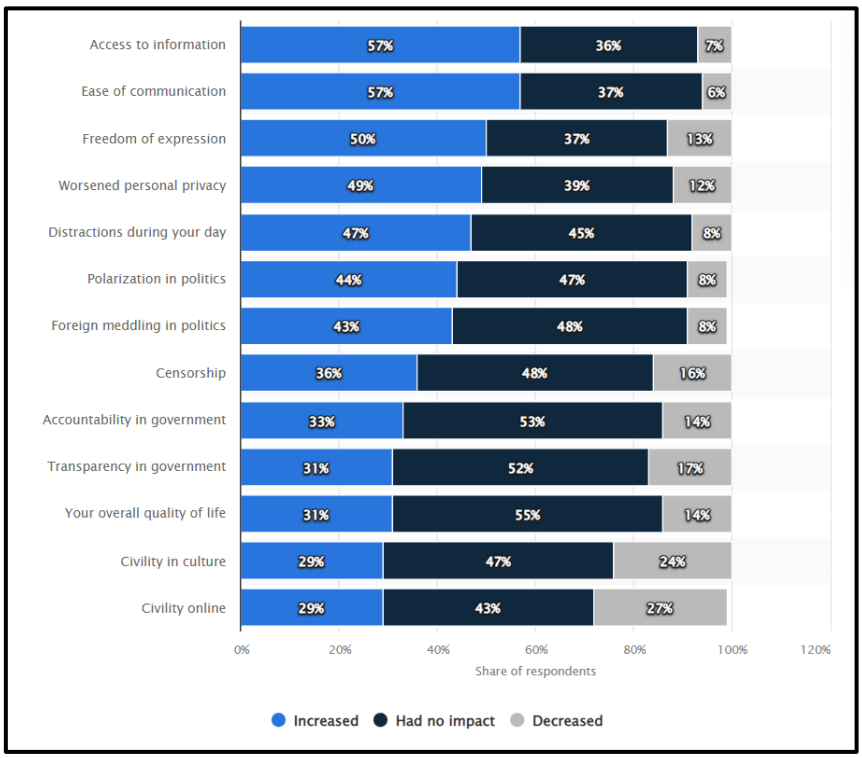


Figure 4.5 Statistics of the use and impact of social media in the daily life worldwide from December 2018 to February 2019

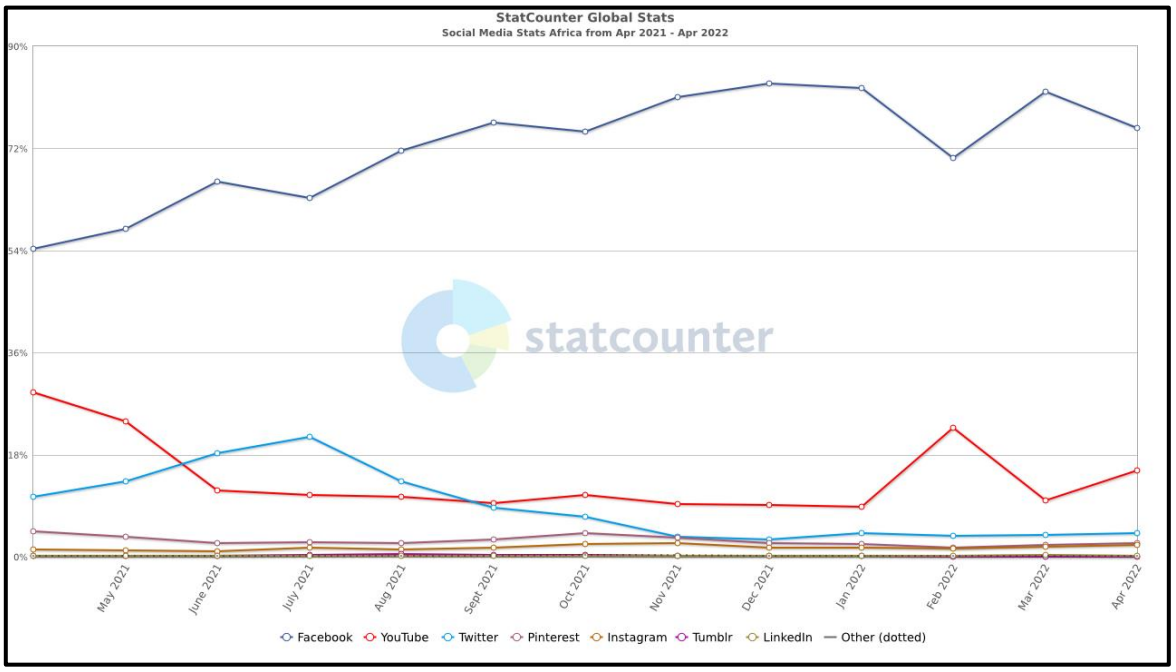


Figure 4.6 Statistics of the use of social media in Africa between May 2021 and April 2022.

According to these statistics, Facebook is the most popular social media and one of the most used mobile applications around the world that has over than 2900 million users in 2022. Thanks to its capability of giving people the chance to access to information they are looking for, to communicate easily with the people they want to, to express freely what they think, and to escape from the real life during the day, Facebook users are growing day by day. Moreover,

Re-searchlight is located in Africa, in which in only one year the use of Facebook has increased from 54% to 82%.

In conclusion, for our research, Facebook might be the best data source as it has the most popularity all over the world and especially in Africa where is most of the company's clients.

4.5.2. Data extracting tool

As we fixed our data source, we will need to specify which tool we are using to extract the desired data.

The most known OSINT tools that we found that works with extracting data from Facebook are: Scrape Storm, SOCMINT tools, Facebook search, Prospectss, Phantom Buster, Apify, Octoparse, Let's extract email studio. The most recommended tools according to the tests we have done are Scrape Storm, Phantom Buster and Let's Extract Email Studio.

The table below presents the difference between these three selected tools.

Table 4.4 Comparison between data collection tools.

Tool name	Tool definition	Advantages	Disadvantages
Scrape Storm	**Scrape Storm** is an AI-Powered visual web-scraping tool that extracts data from almost any website without writing any code. It is powerful and very easy to use.	• Scrape data from Facebook posts. • Results in a CSV file.	• Does not always give the desired results from the scraped posts. • Limited number of data rows even in the paid version. • It is hard to configure the scraping needs.
Phantom Buster	**Phantom Buster** is a technology company that offers data scraping and automation solutions in form of phantoms for over 20 categories online (Facebook, twitter, LinkedIn, etc.).	• Extract post commenters, posts likers, group members, and scrape profiles from Facebook. • Use action limitation to save the Facebook account from getting banned. • Results can be in a csv or JSON file.	• Limited time of use, actions, and number of scraped profiles even in the paid version.

Let's Extract Email Studio	**Let's Extract Email** Studio is a web-based tool that collects an unlimited number of leads.	• Extract group members. • Result can be in the desired form in the paid version.	• Facebook account banned after one use. • The software stops working and the data get lost often.

According to *table 4.4* and the results we have, Phantom Buster might be the perfect tool for our needs as it uses limitations to make the scraping method safer, it extracts all the public information we need, and its results format either JSON or CSV.

4.5.3. Data extraction

The first thing to do in our data collection phase is to create a Google Sheet that contains the target groups and posts as shown in the figure 4.7, figure 4.8 and figure 4.9 .

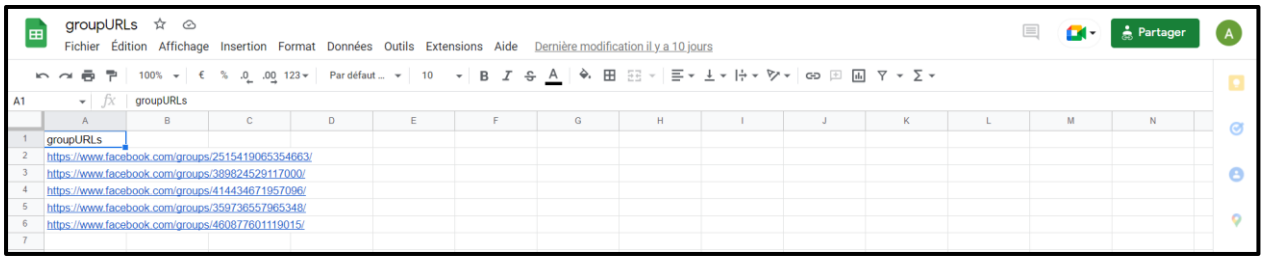


Figure 4.7 Groups URL - Google Sheets.

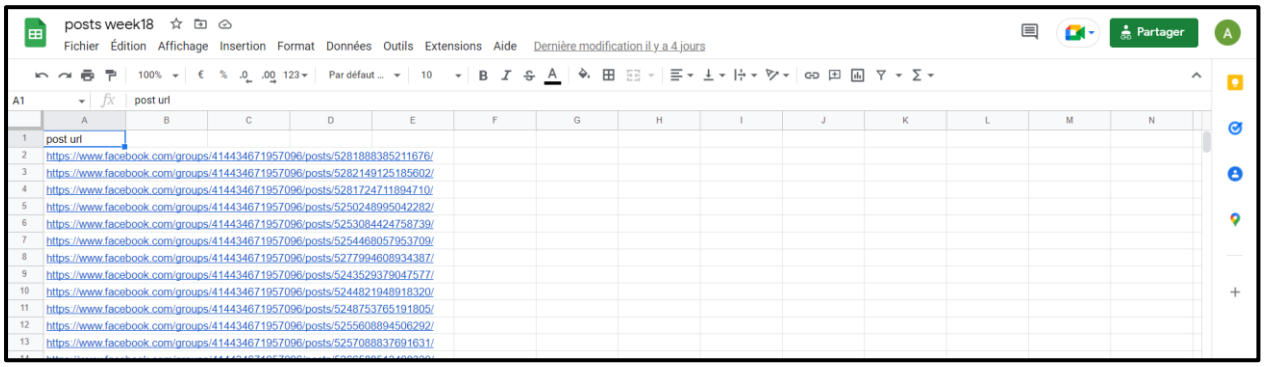


Figure 4.8 Posts week 18 - Google Sheets

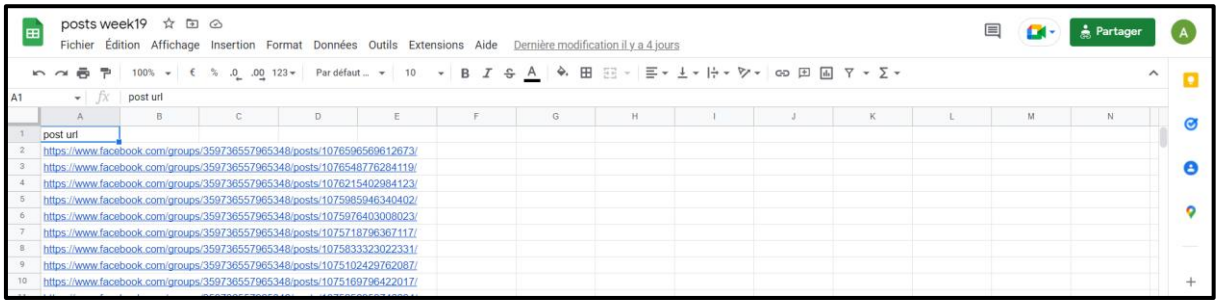


Figure 4.9 Posts week 19 - Google Sheets

For the post's comments and likes, we chose to extract those from the 18th week (from 2 May to 8 May) and the 19th week (from 9 May to 15 May) of 2022. We specified the groups we will work on for these periods, which are "Toutes contre le cancer du seins." and "Denguirat ♡♡ Régime Dr Denguir".

As the Google Sheets are ready, we step now to configure the Phantom Buster phantoms that we will be working with. In our case, we chose four phantoms: Facebook group extractor (to extract Facebook group members), Facebook profile scraper (to scrape Facebook profiles), Facebook post commenters (to extract post commenters), and Facebook post likers (to extract post likers).

The following figures detail the configuration of the phantoms.

- Facebook group extractor and Facebook profile scraper:

Figure 4.10 presents the behavior settings of the Facebook group extractor.

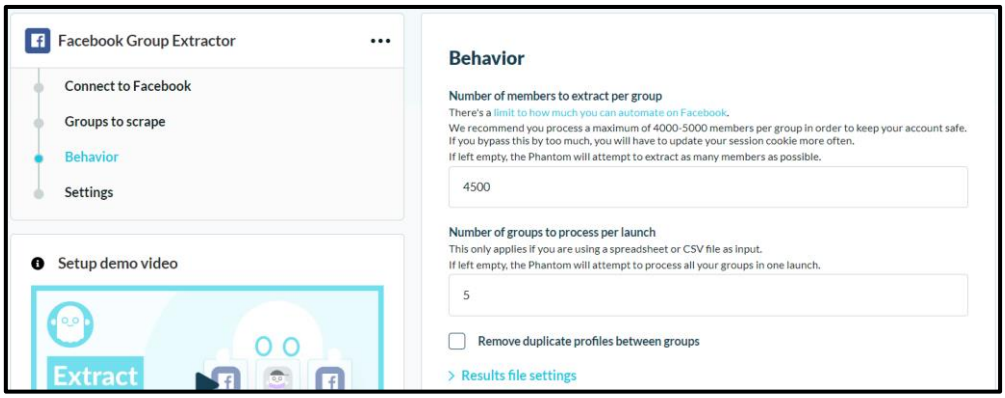


Figure 4.10 Facebook Group Extractor - Behavior

Figure 4.11 presents the launch settings of the Facebook group extractor.

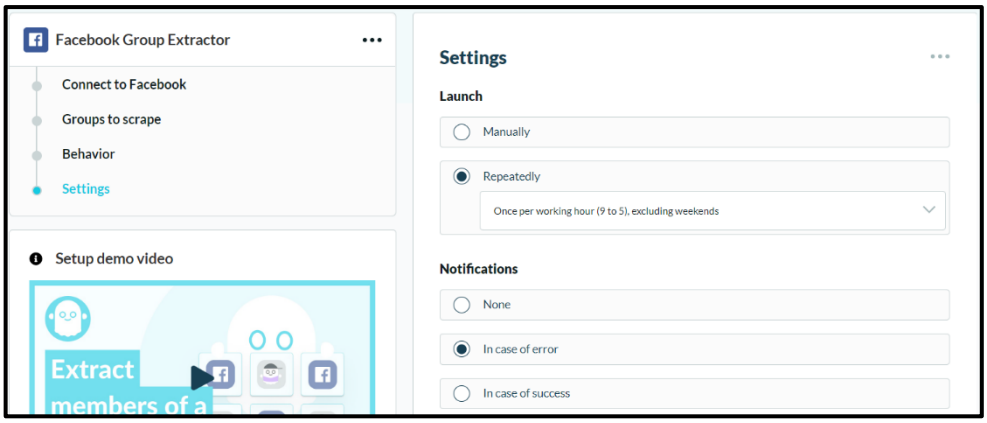


Figure 4.11 Facebook Group Extractor – Settings

Figure 4.12 presents the way Profiles to scrape connects to Facebook group extractor results.

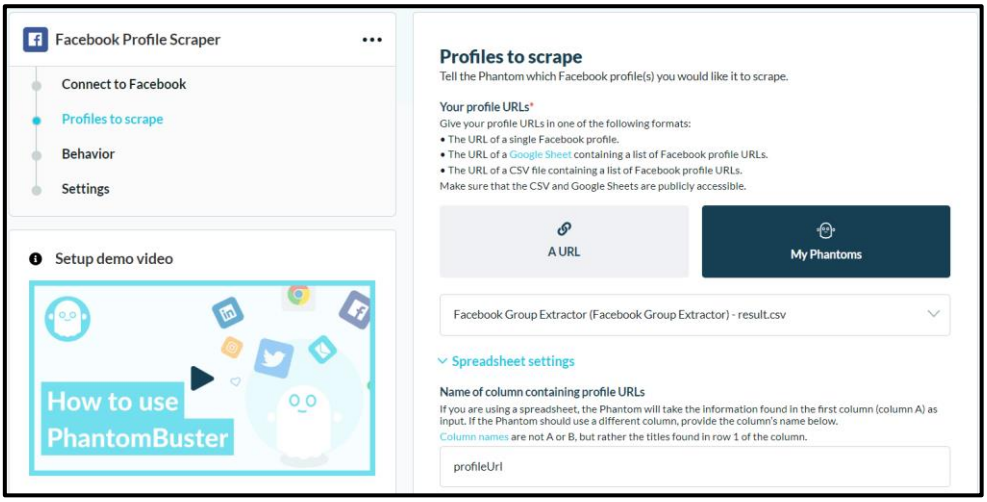


Figure 4.12 Facebook Profile Scraper - Profile to scrape

Figure 4.13 presents the launch settings of the Facebook profile scraper.

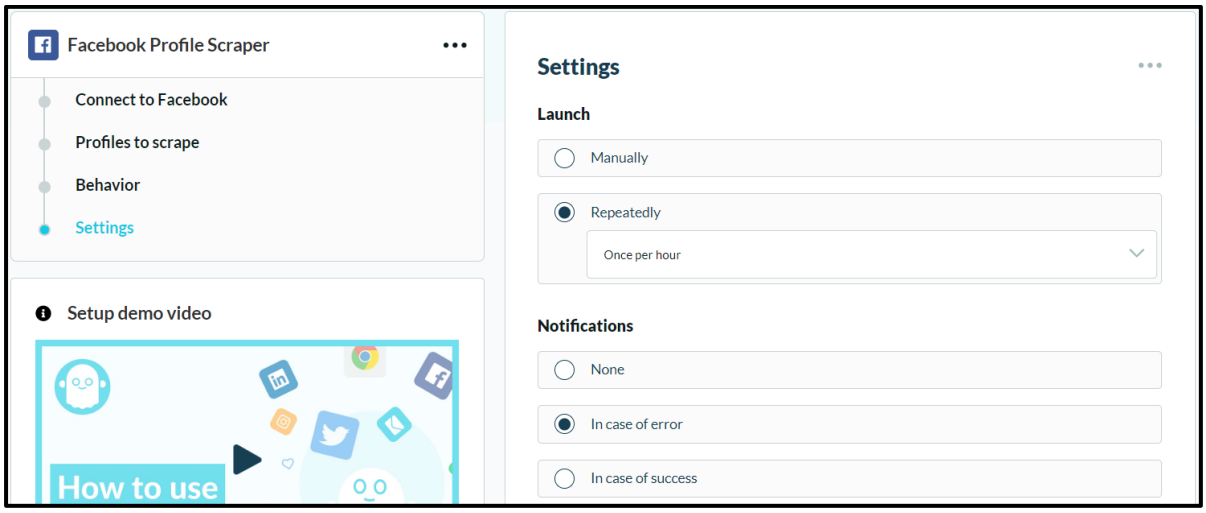


Figure 4.13 Facebook Profile Scraper - Settings

- Facebook post commenters and Facebook post likers:

Figure 4.14 presents the behavior settings of the Facebook post commenters, which is the same for the Facebook post likers.

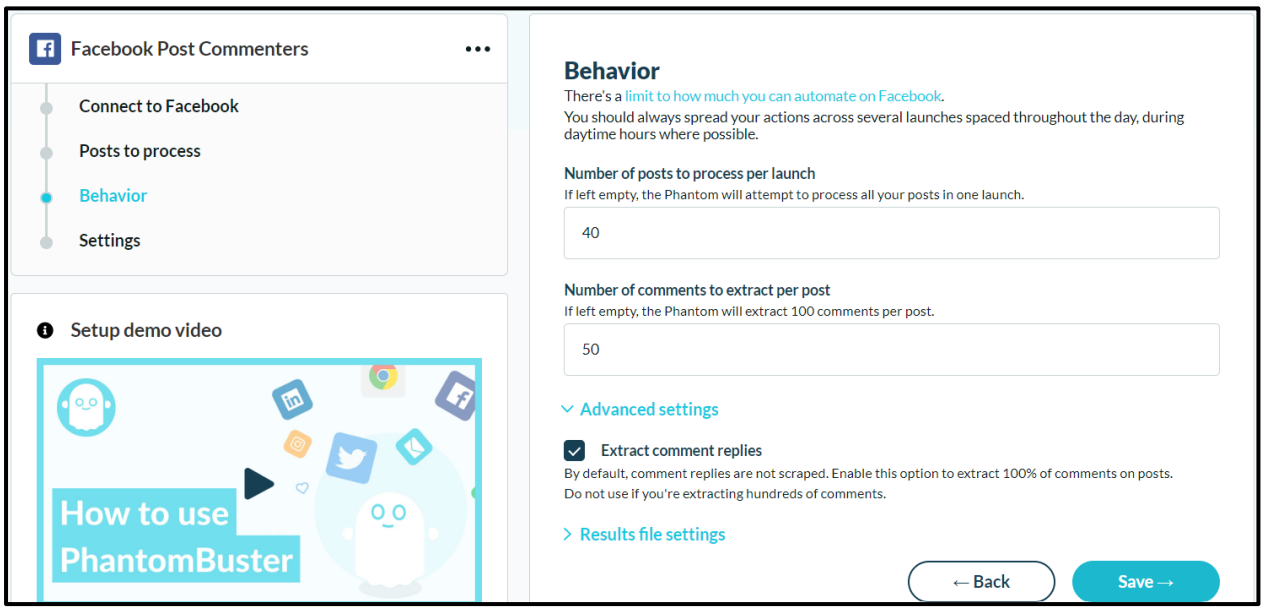


Figure 4.14 Facebook Post Commenters - Behavior

Figure 4.15 presents the way phantom buster presents its phantoms after the configuration and the launch.

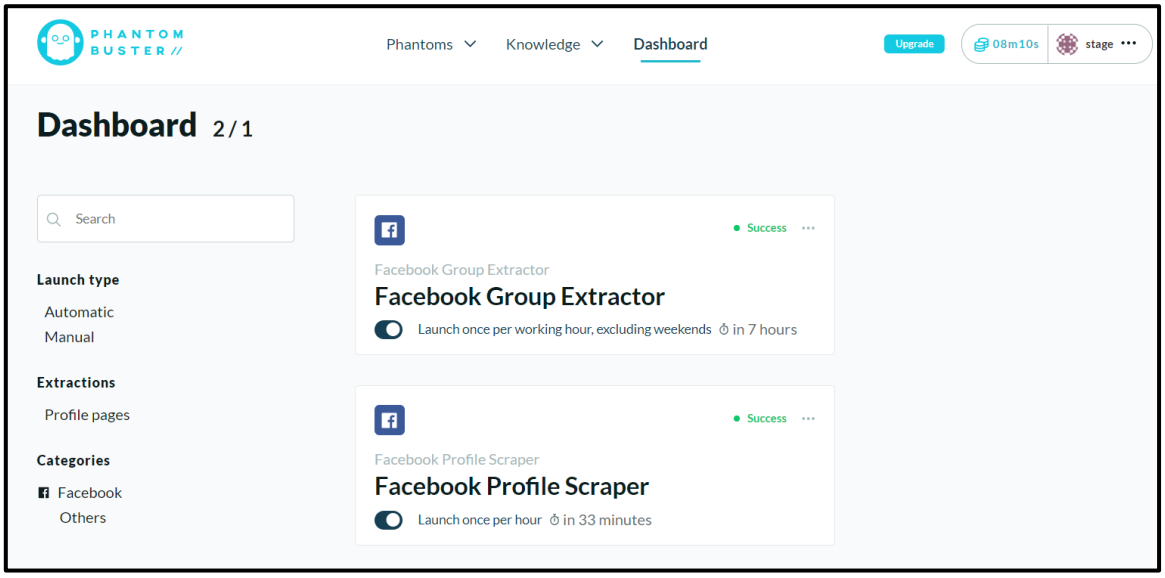


Figure 4.15 Phantom Buster Dashboard

.

When clicking on the phantom, a new page containing the log activity, the results in a JSON or CSV form, and more information about that phantom in the way the following figure illustrates.

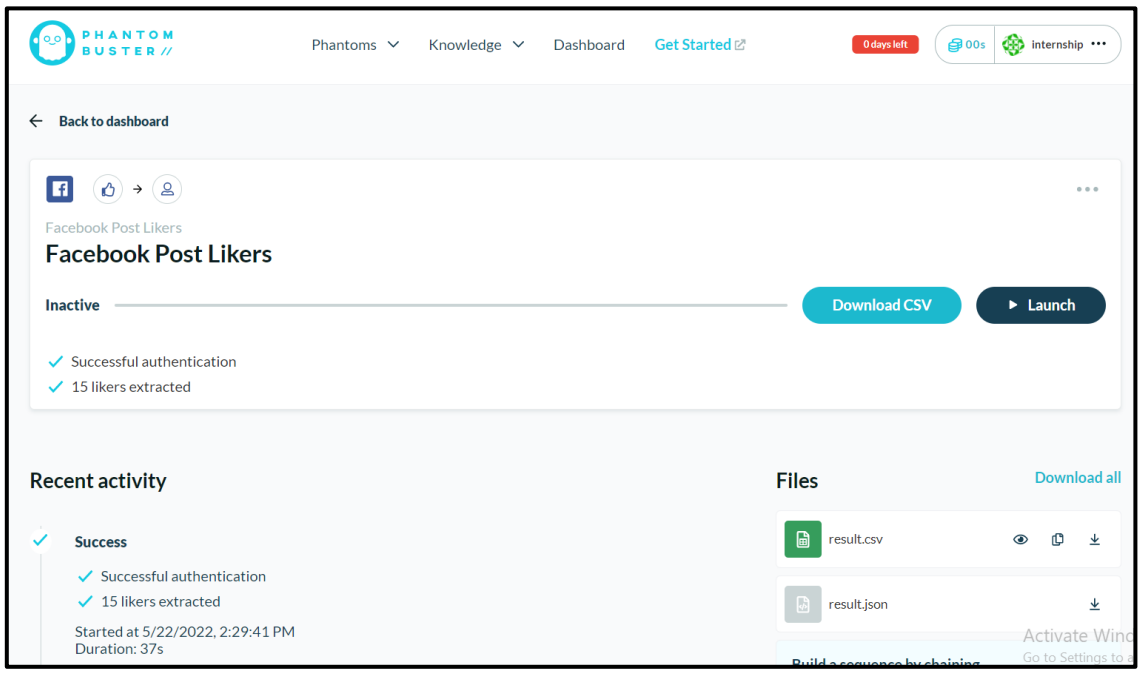


Figure 4.16 Phantom/Agent interface

Phantom Buster gives the user the chance to automate the data extraction, which makes the phantoms launch repeatedly in a defined time interval. The configuration differs from one phantom to another and so do the results.

Table 4.5 shows the results of the phantoms we will use in our process. (Phantom Buster only extracts the publicly available data from a user profile, so some fields might be empty)

Table 4.5 Phantom Buster - phantoms results

Phantom name	**Phantom results**
Facebook group extractor	• Profile URL: the user profile URL • Name: the user name • First Name: the user first name • Last Name: the user last name • Profile Picture: the user profile picture URL • Group Name: the group name in which the user is a member. • Group URL: the group URL • Member Since: since when the user is a member in that group • Additional Data: additional data about the user

	• Friendship Status: can we add this user as a friend or not • User id: the user Facebook id • Timestamp: the system time and date when we extracted the data
Facebook profile scraper	• Full name: the user full name • Education name: the education career of the user • Education URL: links related to the education career • Education Description: the education description • Location Name: the user location, can be more than 1 • Location Type: current or original city of the user • Location URL: links related to the location. • Facebook: Facebook identification • Relationship: relationship status of the user • Relationship URL: relationship URL (partner link) • Profile Picture URL: the user profile picture URL • Banner URL: the Facebook banner URL • Is Friend: can we add this user as a friend or not • Profile URL: the user profile URL • Timestamp: the system time and date when we extracted the data • Work Name: the user work name • Work URL: links related to the work • Work Description: work description (job description/ job period/..) • Gender: user gender • Bio: the user bio • Quote: a quote published by the user on his profile • Instagram: Instagram account • Birthday: date of birth • Interested by: the user interests • Nickname: the user nickname

	• Languages: the user languages • Facebook id: the user Facebook id • Religion: the user's religion • Website: the user website • Phone Number: the user's phone number
Facebook post commenters	• Profile URL: the user profile URL • Name: the user name • First Name: the user's first name • Last Name: the user's last name • Comment: the comment text • Comment Date: the comment published date • Comment Type: the comment type (comment or reply) • Query: the post URL • Timestamp: the system time and date when we extracted the data
Facebook post likers	• Profile URL: the user profile URL • Name: the user name • First Name: the user's first name • Last Name: the user's last name • Profile Picture URL: the user profile picture URL • Can Add Friend: can we add this user as a friend or not • Query: the post URL • Timestamp: the system time and date when we extracted the data

The following figures present a closer look at the results in JSON format.

Figure 4.17 JSON results - part one

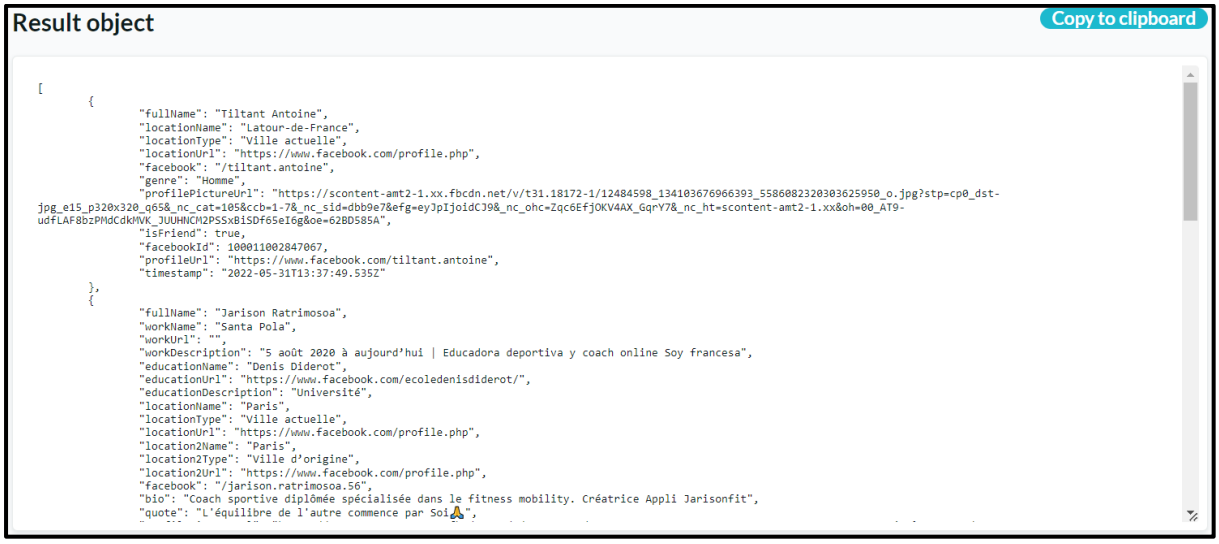


Figure 4.18 JSON results - part two

Figure 4.19 JSON results - part three

4.6. Sprint Review

We present the current sprint's product increments to the Product Owner Mrs. Amel BOURASSI and the scrum master Mrs. Rodile ANNABI during this meeting, which is part of the sprint demonstration and validation process. We approved and validated the features of this first sprint. We will continue with the next sprint.

4.7. Conclusion

We have completed our first sprint at this early stage of our project. We currently have a potentially exploitable first increment of our application.

In the following chapter, we will concentrate on the development of our second sprint, which is relating to data transformation and loading them in a centralized and non-volatile database (*i.e.*, DWH.

CHAPTER 5

SPRINT 2: DATA TRANSFORMATION & LOADING

Chapter 5
Sprint 2: Data Transformation & Loading

5.1. Introduction

After validating the first sprint, we step forward to the second one. This sprint includes a number of features, such as data transformation and data loading.

We begin by developing the sprint backlog. The requirements analysis phase is then presented, which consists of creating the sprint's use case diagram, followed by a textual description of the various functionalities. Following that, the sequence diagram showing the interactions between the actors and the BI system is presented. Finally, we will discuss the realization phase.

5.2. Backlog sprint 2

With the same spirit of the last sprint, we move forward to work on our next sprint, sprint2.

Our goal for this sprint is to complete the last two phases of the ETL process, which are data transforming from messy data to useful information and data loading into the SSMS.

Table 5.1 Sprint Backlog -Sprint 2

ID-US	User story name	User story description	Tasks	Week
2.1	BI Modeler- Create the multi-dimensional model of the DWH	As a BI modeler, I want to create the multi-dimensional model of the DWH in order to load/store useful information in a	Extract Data from social network	W1/W2
			Definition facts and measures using the top-down, bottom-up and mixed approaches	

		centralized and non-volatile database	Definition of dimensions using the top-down, bottom-up and mixed approaches	
3.2	BI Modeler- Transform and clean data	As a BI modeler, I want to transform and clean data in order to convert the data into useful information.	Convert data to a standardized format	W3/W4
			Combine data source (*e.g.*, by a precise mapping to key values)	
			Application of filters	
3.3	BI Modeler- Store data in a DWH	As a BI modeler, I want to load/store useful information in a centralized and non-volatile database (i.e., DWH).	Define the procedures for loading data.	W5/W6
			Feed the tables.	

5.3. Requirements analysis

We identified the sprint's actor and described the various features/user stories with their use cases during this step.

5.3.1. Use case diagram

The use case diagram "sprint2- data transformation and loading" is presented below in figure 5.1, along with a textual description of each case/functionality.

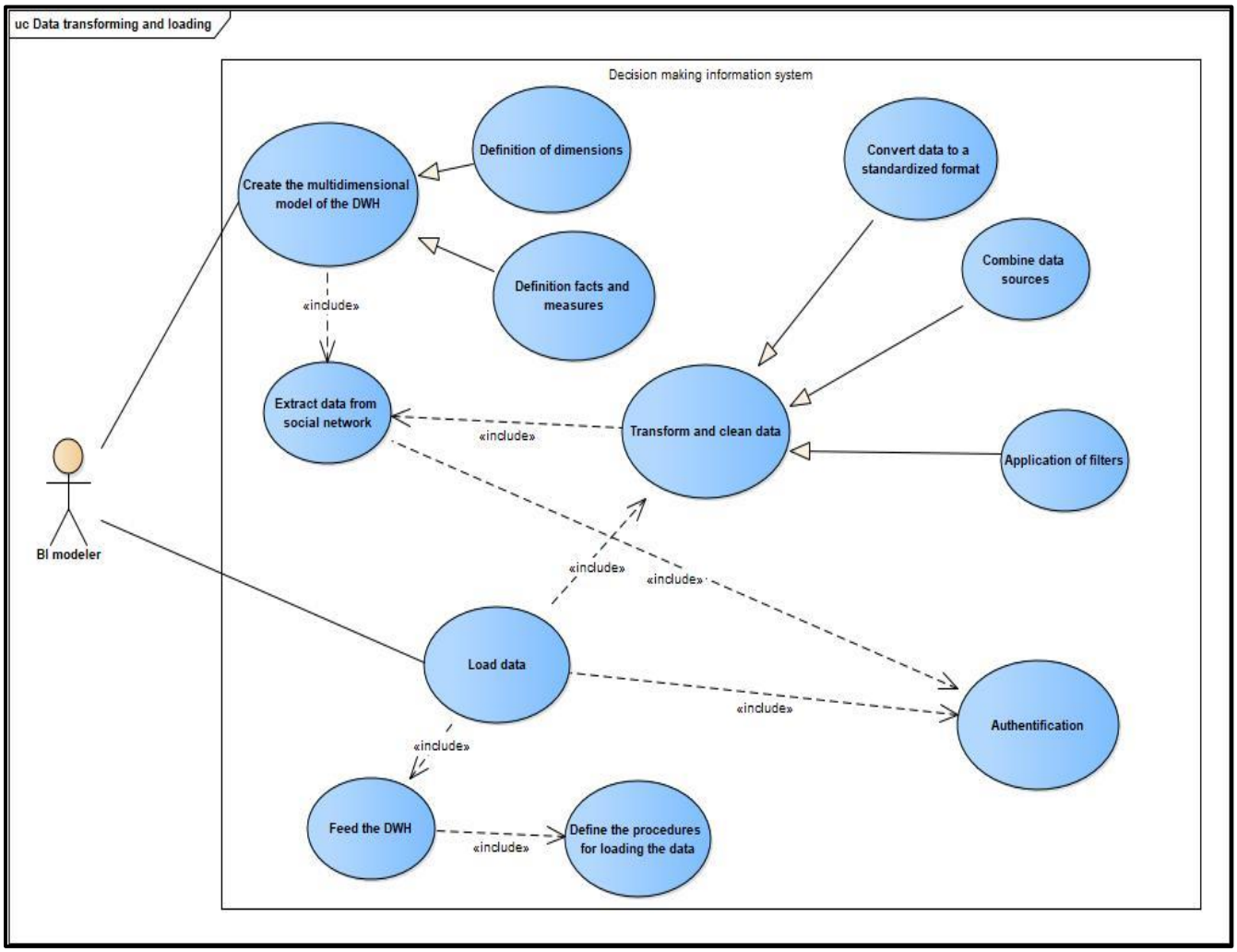


Figure 5.1 Use case diagram-Sprint 2

5.3.2. Textual description of use cases

Table 5.2 Textual description of the use case "Data transform and load"

Title	**Data transform and load**
Resume	The system allows the BI modeler to clear data and load its useful information into a data warehouse.
Actor	BI modeler
Precondition	The BI Modeler has the right to access the extracted data. The BI Modeler has the right to access the data warehouse.
Nominal Scenario	BI Modeler access to the extracted data. BI Modeler transforms and clears the data.

	BI Modeler determines the fact, dimensions, and measures. BI Modeler creates the multidimensional schema of the DWH. BI Modeler defines the procedure for loading data. BI Modeler connects to the data warehouse. BI Modeler loads the data into the DWH.
Post-condition	Clear data is inserted into a data warehouse.
Complements	BI Modeler must determine the necessary data to work with in order to satisfy the client's requirements.

5.4. Conception

The sequence diagram and the class diagram are used to translate conceptual modeling, which is the second activity in a sprint. In this work, we used a sequence diagram to represent the interactions between the actors and the system in chronological order.

5.4.1. Sequence diagram

In order to clear and save the extracted data, the system executes a script that asks to get the JSON results from the Phantom Buster and save them in the PostgreSQL database as an intermediate step, in order to facilitate the work process.

Next, the system executes another script that converts the data into a standardized format and starts cleaning the data rows.

The system will ask to reinsert the data into another PostgreSQL table, so it can ask talend to continue the transformation process.

Talend requests the data rows from PostgreSQL to continue the data transformation step.

The system demand talend to save the data. Talend saves the data in SSMS and returns a success message to the system.

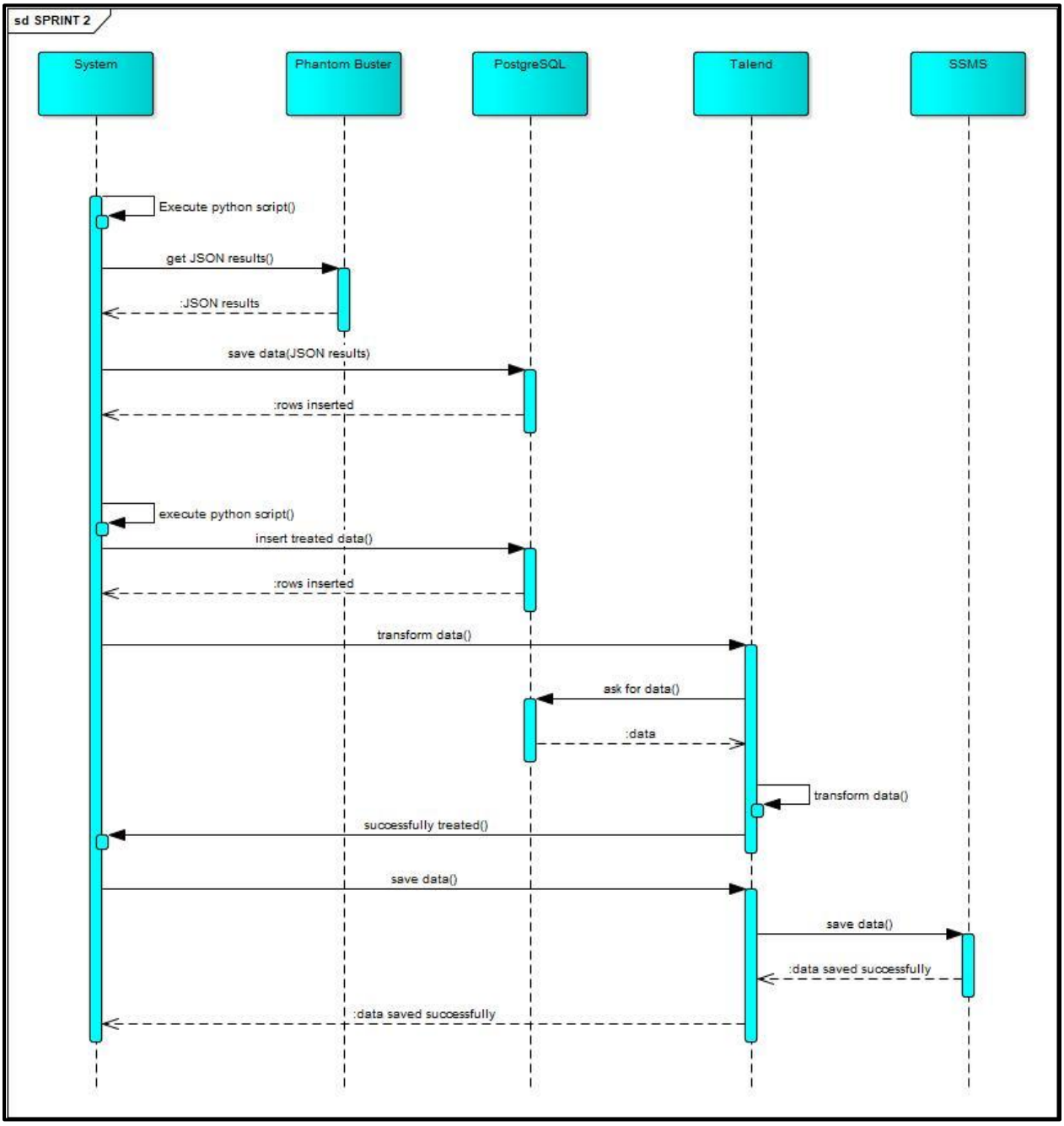


Figure 5.2 Sequence diagram- Sprint 2

5.5. Realization

5.5.1. Data transformation (T)

After extracting the data from the desired sources, we need the result object part from the JSON results to be able to start the transform step. We chose to save this data in the PostgreSQL database

as an intermediate step to facilitate the work because Python supports the PostgreSQL database adapter and the Phantom Buster API.

- **The Phantom Buster API**: gives us control over our account. It is composed of HTTPS endpoints returning JSON data

We used the following libraries with python:

- **Requests**: Python library for making HTTP requests so that we can focus on interacting with services and on consuming data in our application.

  In our case, we needed:

  - Our account API Key: to identify with which Phantom Buster we are trying to connect. It is a unique key for each account.
  - The agent ID: that presents the phantom ID.

  Figure 5.3 presents the way we used this requests library in our Python script.

Figure 5.3 Phantom Buster connection using Phantom Buster API in python script

- **Psycopg2**: Python library used as a PostgreSQL database adapter. This library allows us to connect to our database and make changes like the CRUD (CREATE, READ, UPDATE, DELETE) operations.

  The following figure presents the psycopg2 library logo.

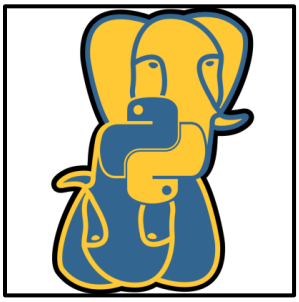


Figure 5.4 Psycopg2 python library logo

We used the psycopg2 to connect to PostgreSQL and insert the data rows in the database as shows figure 5.5.

members.py

```
# instantiate a cursor object from the connection
# declare a new PostgreSQL connection object
conn = connect (
        dbname = "Stage2022",
        user = "postgres",
        host = "localhost",
        port = "5432",
        password = "st!ge!22"
    )

# attempt to create a cursor object
cursor = conn.cursor()

# check if the connection was valid
if cursor != None:
    print ("\nconnection successful:", conn, "\n")

sql_string = """
INSERT INTO members SELECT *
FROM json_populate_recordset(NULL::members, %s);
"""
json_result = json.dumps(json_result)
cursor.execute(sql_string, ([json_result,]))
conn.commit()

# close the cursor object to avoid memory leaks
cursor.close()

# close the connection object
conn.close()
print ("\nconnection closed:")
```


Figure 5.5 PostgreSQL database connection using python

We repeated the process for the following tables: comments, likes, profilescrape, and members.

The result of saving the data was something like the following figure shows.

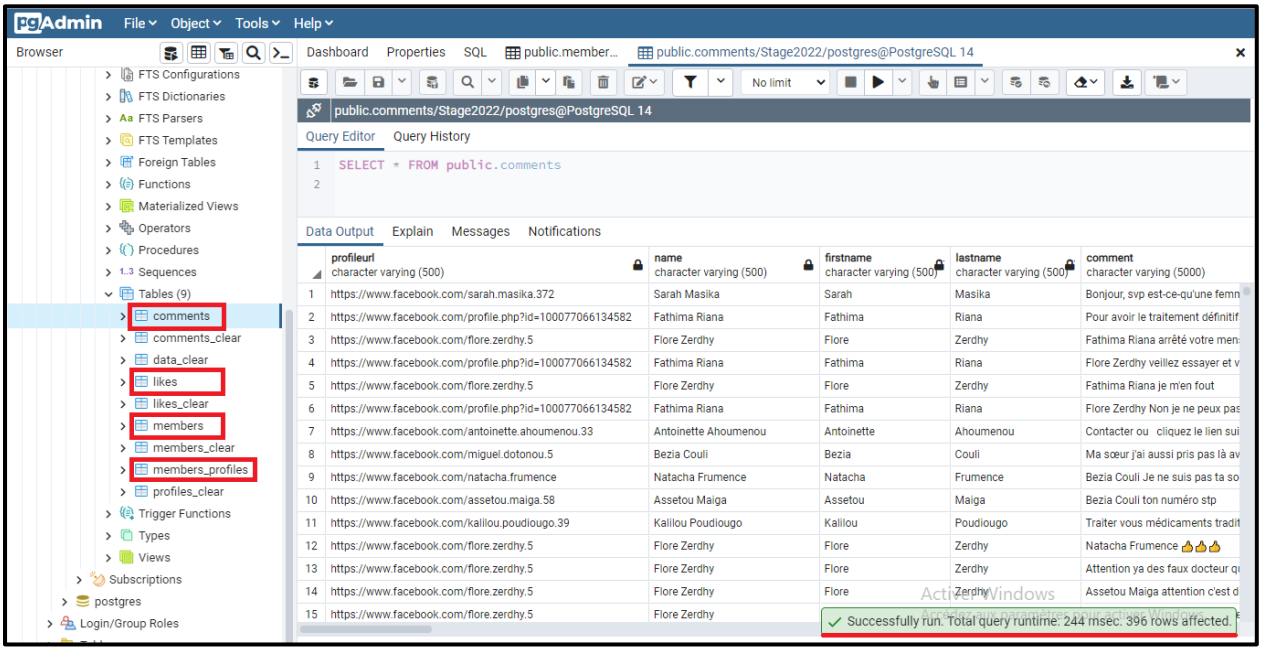


Figure 5.6 PgAdmin interface - comments table

From the collected data, we found that there are many redundant and non-useful data, so we performed a cleaning step to eliminate and reformat the non-useful data and avoid duplicates and other inconsistencies.

For the data cleaning and transformation, we used a script python that we wrote and the "Talend for Data integration" tool. The following figures present some Talend workspace screenshots.

This figure below presents the user profile job, in which we managed to use

- The **tDBConnection** component: to connect to the PostgreSQL database.
- The **tDBInput** component: to get the database table data.
- The **tMap** component: to merge the data tables and treat its columns.
- The **tUniqueRow** component: to remove duplicates.
- The **tSortRow** component: to sort rows.
- The **tDBCommit** component: to validate the data processed through the Job into the connected database.
- The **tDBOutput** component: to send the output to PostgreSQL database.

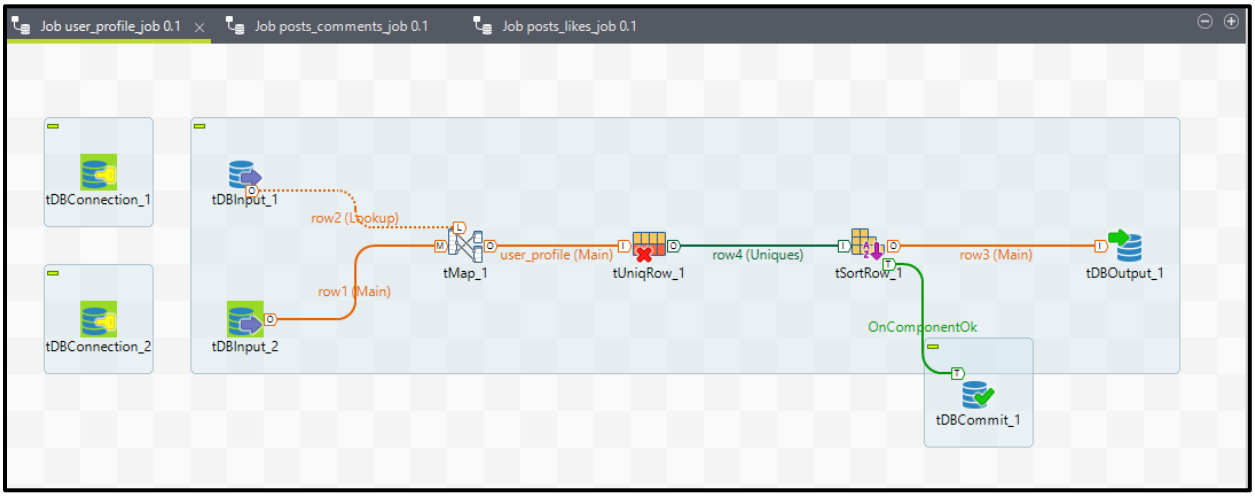


Figure 5.7 Talend workspace - user profile job

The following figures presents the way **tMap** is configured in our system.

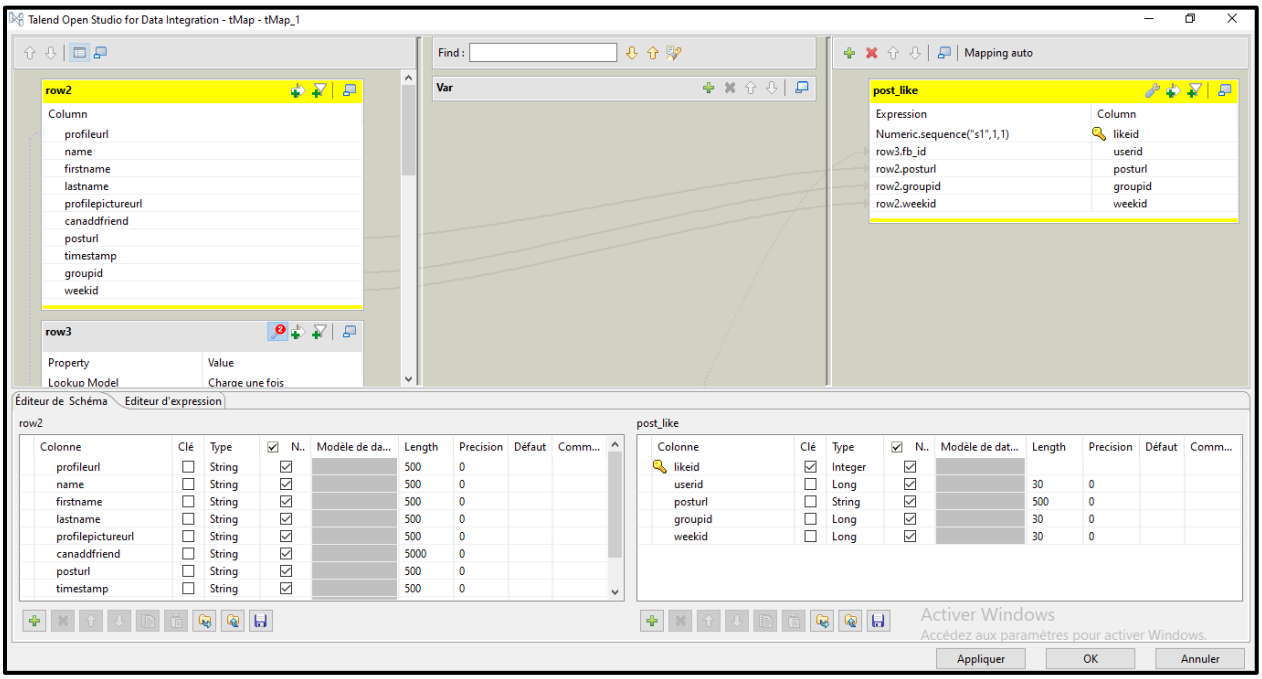


Figure 5.8 Talend tMap component - Like

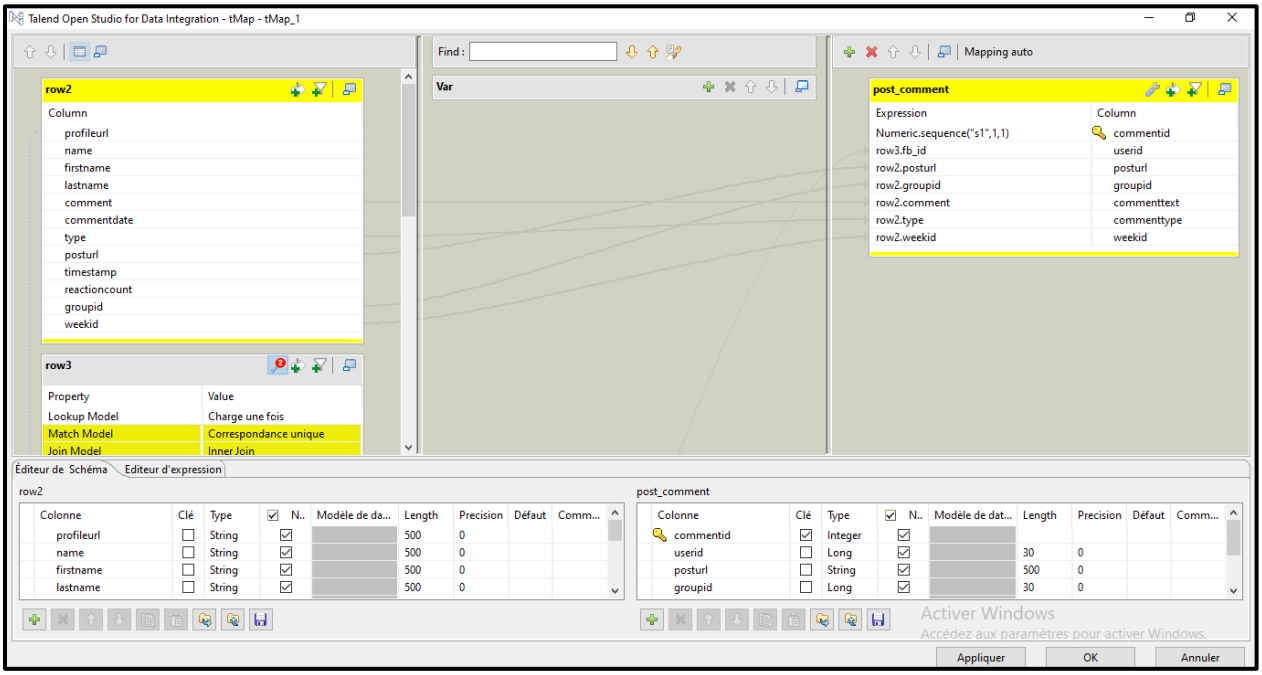


Figure 5.9 Talend tMap component - Comment

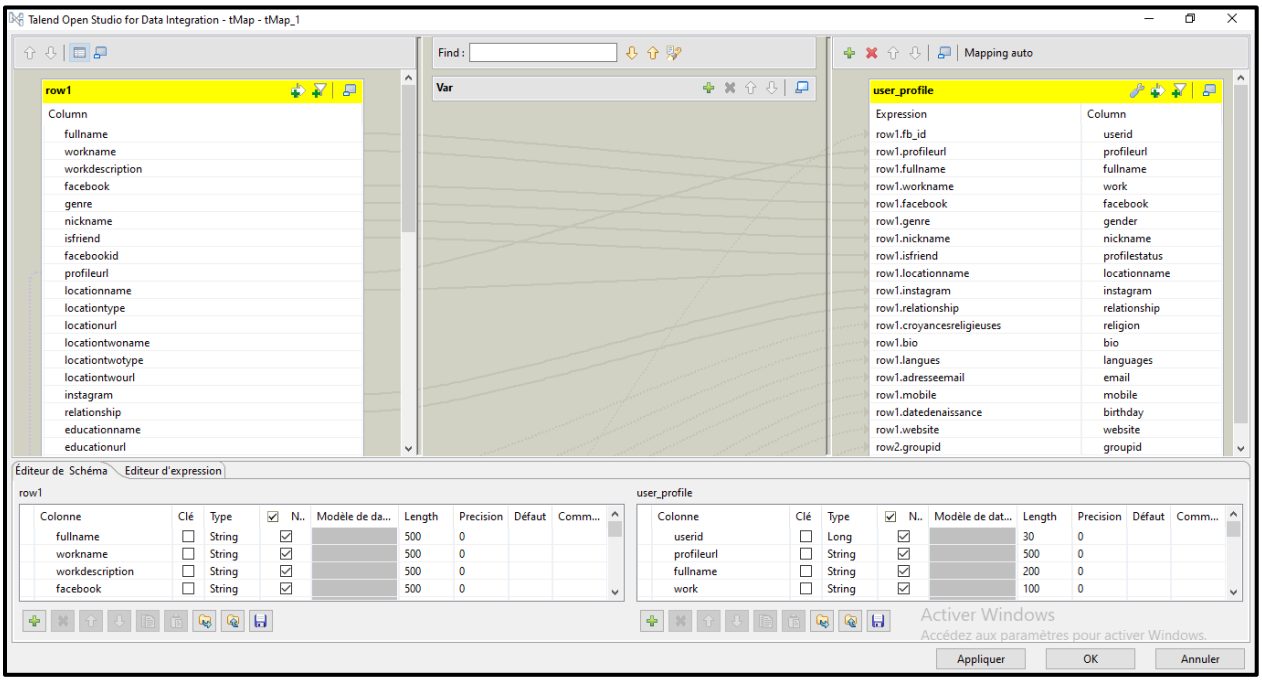


Figure 5.10 Talend tMap component - User profile

The figure below presents an example of transformed data.

Data Output | Explain | Messages | Notifications | Query Editor | Query History

	userid [PK] bigint	profileurl character varying (500)	fullname character varying (200)	work character varying (100)	facebook character varying (100)	gender character varying (100)
1	1128422619	https://www.facebook.com/amandine.marnette	Amandine Marnette	Student	/amandine.marnette	Femme
2	1280213509	https://www.facebook.com/gitanyazid	Yazid Gaci	Not specified	/gitanyazid	Homme
3	1438955537	https://www.facebook.com/maeva.barthod	Maëva Hahaa	Student	/maeva.barthod	Femme
4	1498715635	https://www.facebook.com/gaelle.rougeot	Gaelle Rougeot	Student	/gaelle.rougeot	Femme
5	1508305192	https://www.facebook.com/marie.do.9404	Marie Do	Student	/marie.do.9404	Femme
6	1526191808	https://www.facebook.com/hani.ptth	Hani Btt Manutea	Student	/hani.ptth	
7	1644597161	https://www.facebook.com/hadjer.haned	هاجر حاند	Student	/hadjer.haned	Femme
8	100000241728661	https://www.facebook.com/oceana.linda	Linda Fourchaud	Employee	/oceana.linda	Femme
9	100000984671347	https://www.facebook.com/verohanta.razafiarisoa	Vero Hanta Razafiarisoa	Not specified	/verohanta.razafiarisoa	Femme
10	100001390363385	https://www.facebook.com/sandrine.shinakawa	Sandrine La Shinakawa Bondo	Employee	/sandrine.shinakawa	Femme
11	100002080959225	https://www.facebook.com/sdirichadia	Chadia Jendoubi	Student	/sdirichadia	
12	100003547916575	https://www.facebook.com/karybacha	Rima Hacene	Student	/karybacha	Femme
13	100004259081434	https://www.facebook.com/blanche.charreton	Blanche Charreton	Employee	/blanche.charreton	Femme
14	100004753061001	https://www.facebook.com/befresh.moramedhusseinosama	Youvedor Dieumaitre	Not specified	/befresh.moramedhusseinosama	Homme
15	100004846839402	https://www.facebook.com/fabienne.cossart.3	Fabienne Baillard	Student	/fabienne.cossart.3	Femme
16	100005705538775	https://www.facebook.com/profile.php?id=100005705538775	Naïmatou Zeba	Student		Femme
17	100005705918524	https://www.facebook.com/bizabishaka.leila	Bizabishaka Leila	Student	/bizabishaka.leila	Femme
18	100005895245711	https://www.facebook.com/marie.timmermans.902	Marie Claude Mathieu	Employee	/marie.timmermans.902	Femme
19	100006679935539	https://www.facebook.com/profile.php?id=100006679935539	Nicole Plante	Student		Femme
20	100007965661510	https://www.facebook.com/emilie.bombaka	Anne Prophétesse Bombaka	Employee	/emilie.bombaka	Femme

Figure 5.11 Clean data example

5.5.2. Data source class diagram

Before we start designing our multidimensional schema, we present the class diagram of our data source. This schema was defined from the PostgreSQL database, which presents the different information collected from Facebook. An analysis of these tables allowed us to establish the following class diagram in order to apply our bottom-up approach. Our source is composed of five tables that are linked together by a composition link as shown in the figure 5.12.

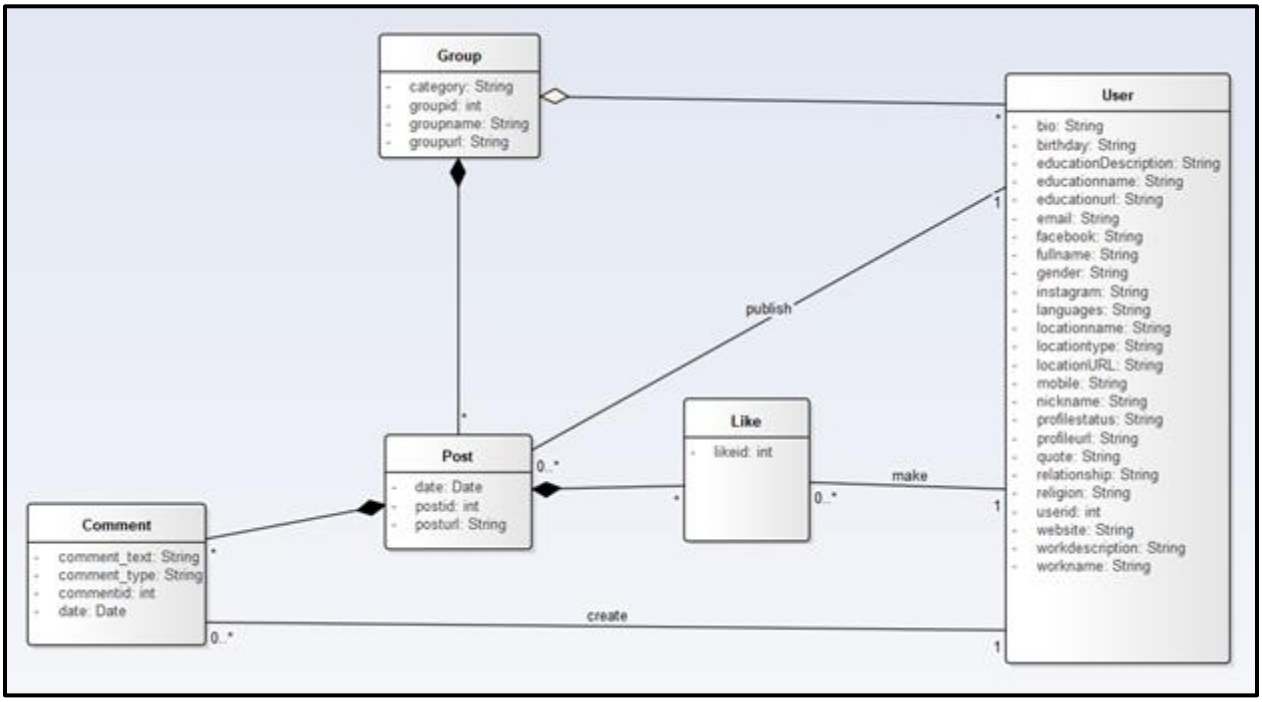


Figure 5.12 Data source class diagram

5.5.3. Multidimensional concepts

This part consists in defining the multidimensional concepts: fact, measure, dimension and hierarchy from the data collected and cleaned from our target Facebook groups.

There are three methods to bring all the elements together to implement a data warehouse, which is as follows:

5.5.3.1. Top-down approach

- **Data collection**

R1: **analyze** the number of comments **per** user **per** group **for** week 18.

R2: **analyze** the number of likes **per** group **for** week 19.

R3: **analyze** the user profiles **per** category.

R4: **analyze** the user profiles **per** gender **per** group **for** week 18.

- **Specification of needs**

Table 5.3 Specification of needs - top down approach

	userid	groupid	weekid	category	gender
nbComment	X	X	X	X	X
nbLike	X	X	X	X	X
profiles	X	X	X	X	X

- **Formalization of needs**
  - **Facts**

    F1: factLike {nbLike (sum, avg, max, min)}

    F1: factComment {nbComment (sum, avg, max, min)}
  - **Dimensions**

    D1: dimUser {userid, fullname, profileurl, gender}

    D2: dimGroup {groupid, groupname, groupurl, category}

    D3: dimTime {iddate, week, year}
- **Multidimensional schema**

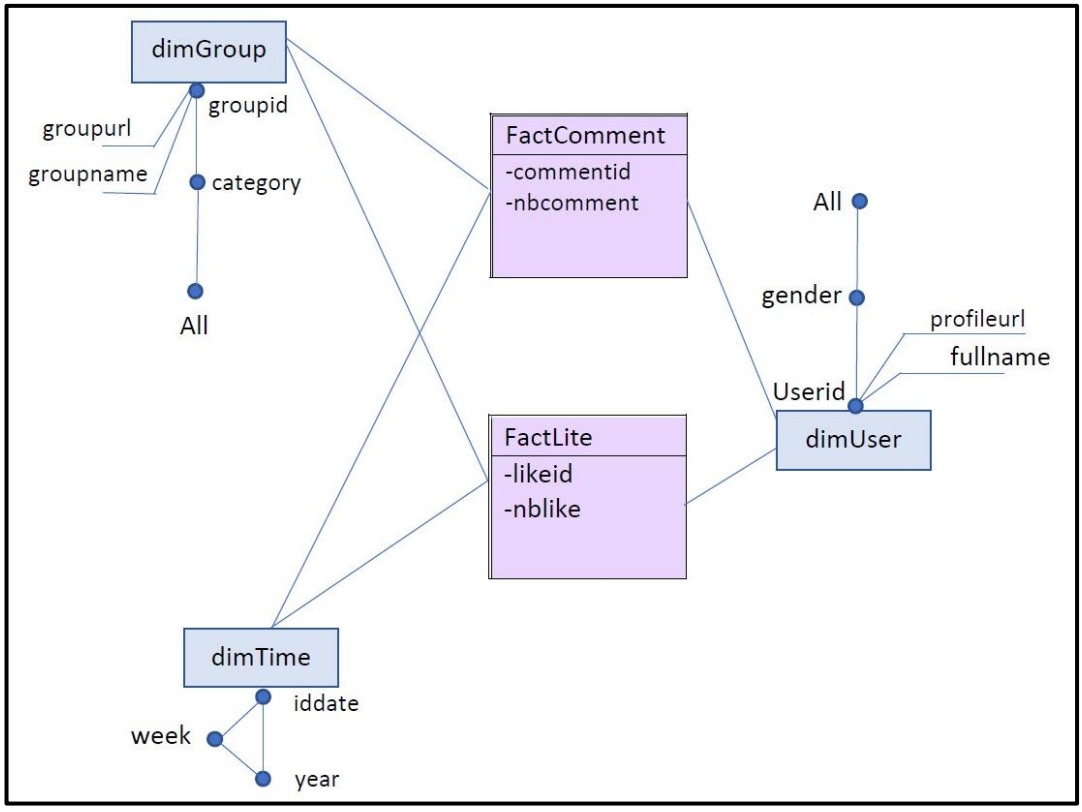


Figure 5.13 Multidimensional schema - top down approach

5.5.3.2. Bottom-up approach

- **Determination of facts:**

This step consists in detecting the representative classes of the analysis. A representative class (RC) describes an event that occurs at a given time and it contains the analysis measures.

F1: factComment {commenteid = S.Comment.commentid, nbComment (sum, avg, max, min)}

F2: factLike {likeid = S.Like.likeid, nbLike (sum, avg, max, min)}

- **Determination of dimensions:**

D1: dimUser {userid = S.User.userid, fullname = S.User.fullname, profileurl = S.User.profileurl, gender = S.User.gender, workname = S.User.workname, facebook = S.User.facebook, nickname = S.User.nickname, profilestatus = S.User.profilestatus, locationname = S.User.locationname, Instagram = S.User.instagram, relationship = S.User.relationship, religion = S.User.religion, bio = S.User.bio, languages = S.User.languages, email = S.User.email, mobile = S.User.mobile, birthday = S.User.birthday, website = S.User.website}

D2: di mGroup {groupid = S.Group.groupid, groupname = S.Group.groupname, groupurl = S.Group.groupurl, category = S.Group.category}

D3: dimTime {idweek, weekdesc}

- **Hierarchization of dimensions and definition of the granularity of the analysis:**

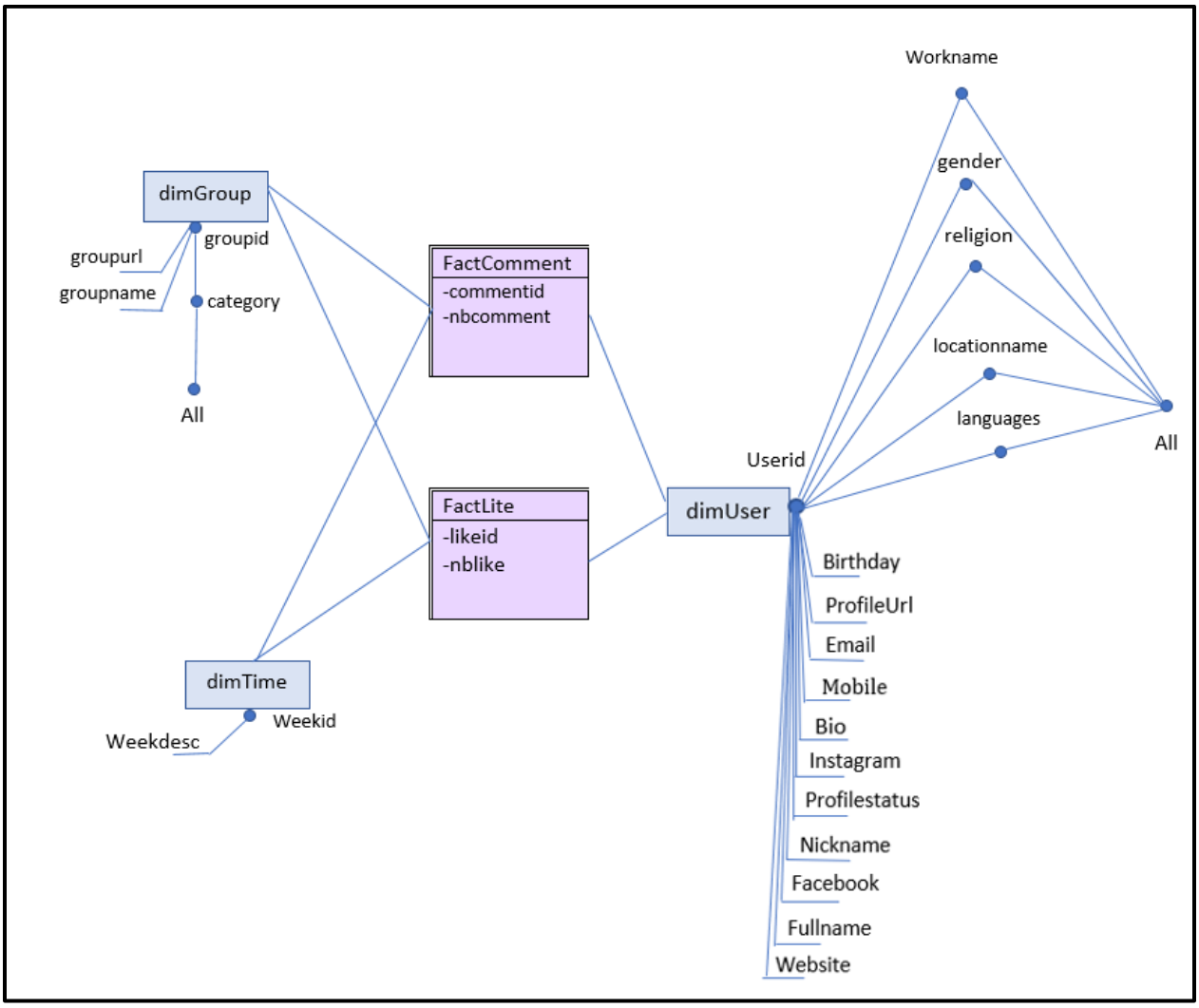


Figure 5.14 Multidimensional schema - bottom up approach

5.5.3.3. Mixed approach

It is a hybrid approach, which combines bottom-up and top-down approaches. In fact, it takes into consideration the data sources and the users' needs.

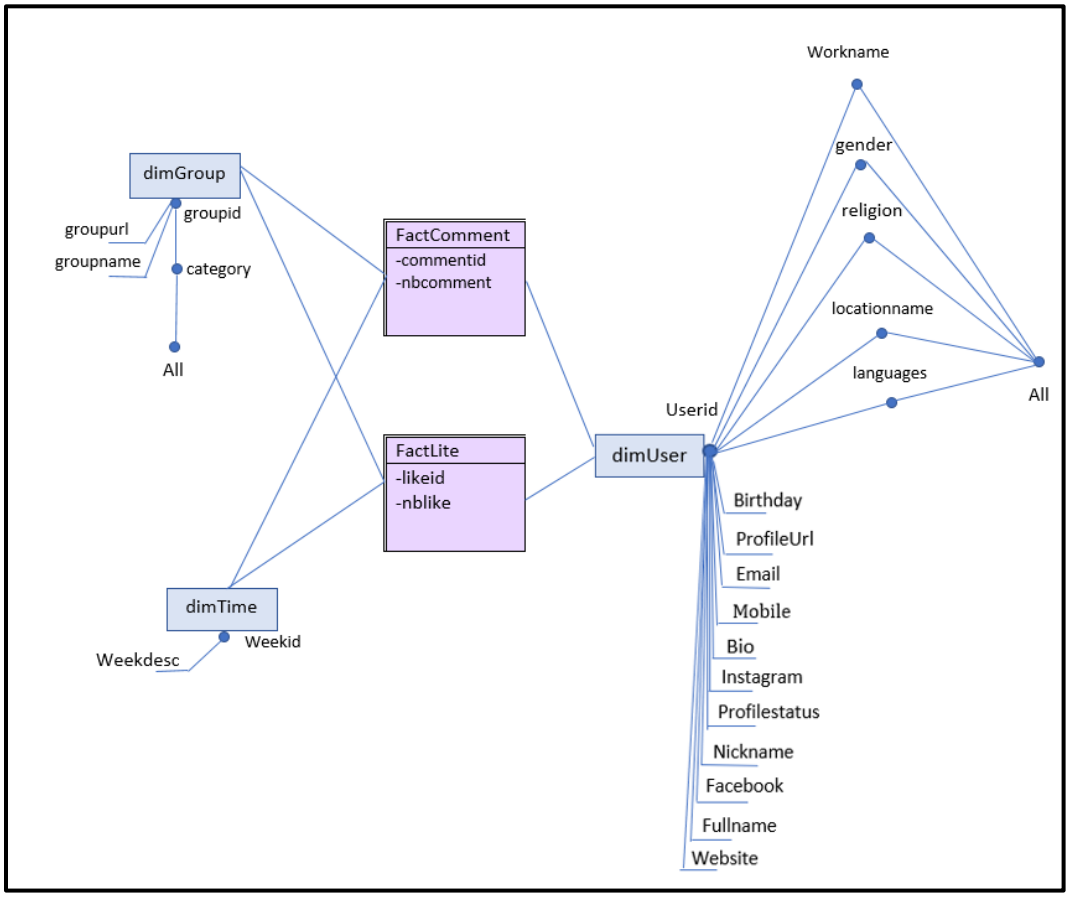


Figure 5.15 Multidimensional schema - Mixed approach schema

5.5.4. Data loading

After defining the schema of the data warehouse in the previous paragraph. We have chosen to model our warehouse using the fact constellation model.

The following figure presents our data warehouse diagram.

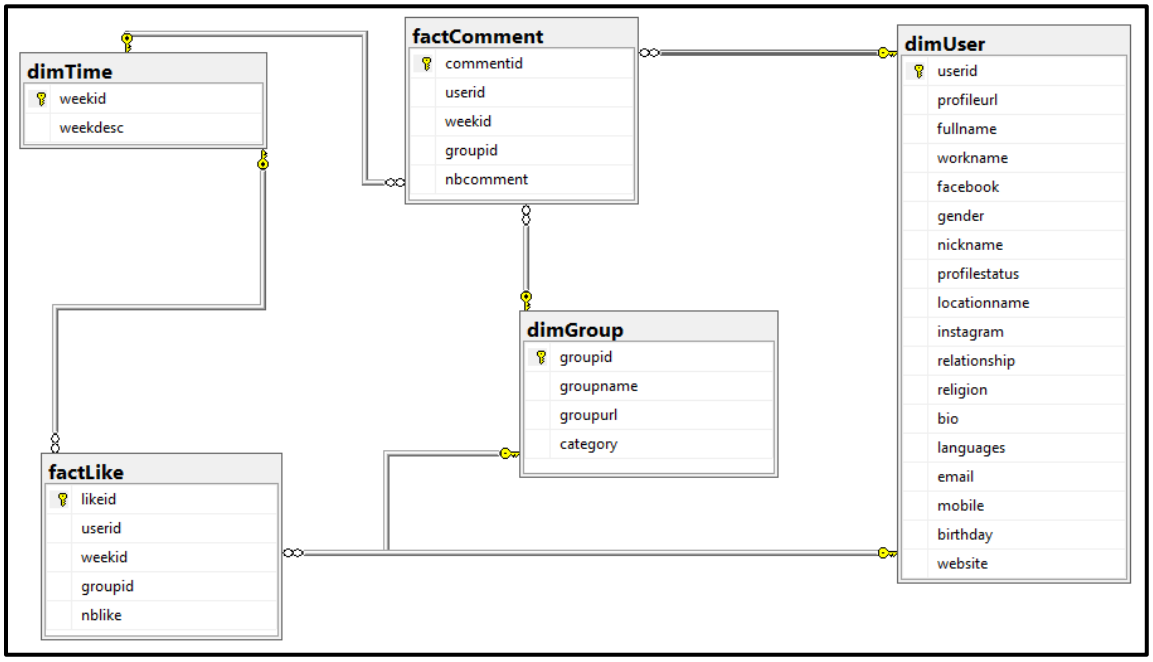


Figure 5.16 DWH diagram

Now that we have created our dimension tables and fact tables, we used Talend to send the data to Microsoft SQL Server.

The following figure presents a job example of filling the data warehouse “dimUser” table.

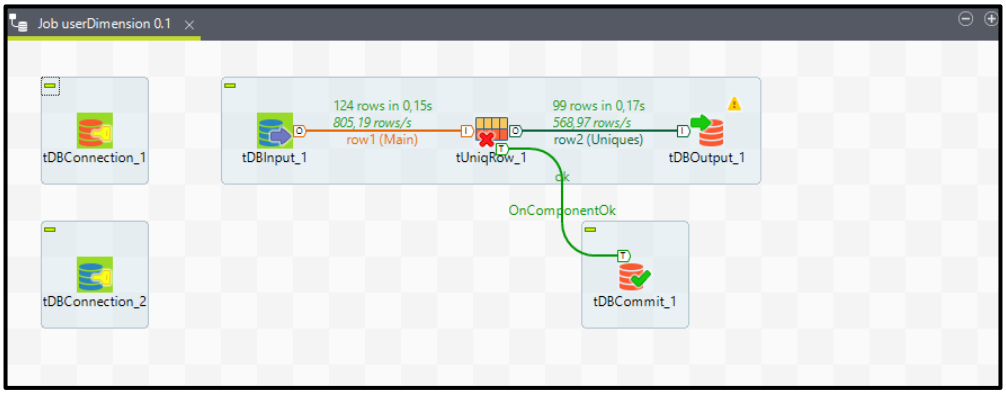


Figure 5.17 Data loading with Talend job

The following figure presents the way our dimUser is filled with clean data.

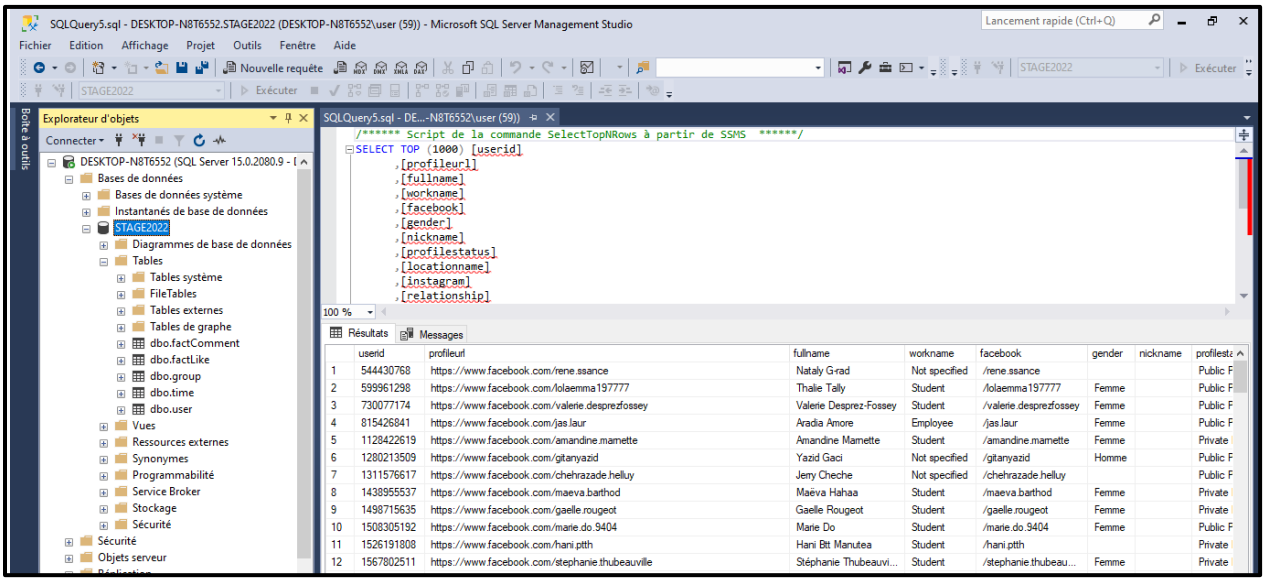


Figure 5.18 SSMS interface - userDim table

5.6. Sprint review

At the end of the sprint, we had a retrospective meeting with our scrum master Mrs. Amel BOURASSI and our product owner Mrs. Rodile ANNABI. After testing our system, they confirmed the features of this sprint.

5.7. Conclusion

In this chapter, we have shown the backlog of the sprint "data transformation and loading" which describes the processed tasks.

Now all that is left is to visualize the data on a dashboard, which will be our goal for the next sprint.

CHAPTER 6
SPRINT 3: DATA VISUALIZATION

Chapter 6
Sprint 3: Data Visualization

6.1. Introduction

We proceed to the final sprint after validating the second one. The data visualization phase is included in this sprint.

We begin, as in previous sprints, by creating the sprint backlog. The requirements analysis phase is then presented, which consists of creating a use case diagram for the sprint, followed by a textual description of the various functionalities. The sequence diagram depicting the interactions between the actors and the BI system is then presented. Finally, we will go over the implementation phase.

6.2. Backlog sprint 3

With the same spirit of the last sprint, we move forward to work on our next sprint, sprint3.

Our goal for this sprint is to complete the last phases of the BI process, which is data visualization using the power BI desktop.

Table 6.1 Sprint Backlog -Sprint 3

ID-US	User story name	User story description	Tasks	Week
4.1	BI analyst-Visualize/analyze data/relevant information	As a data analyst, I want to have real-time data.	Choose a measure and one or more dimensions to display the desired data.	W1/W2
			Start the query.	
			The system returns the result of the query.	
4.2	Configure reports		Create reports	W2

		As a data analyst, I want to configure reports depending on the company's needs.	Consult reports	
5	Make business decisions	As a business manager, I want to easily access the dashboard and consult the visual reports in order to make business decisions.	Create reports Consult reports	W2
6	Make marketing strategies	As a marketing manager, I want to easily access the dashboard and consult the visual reports in order to make marketing strategies	Consult reports	W2

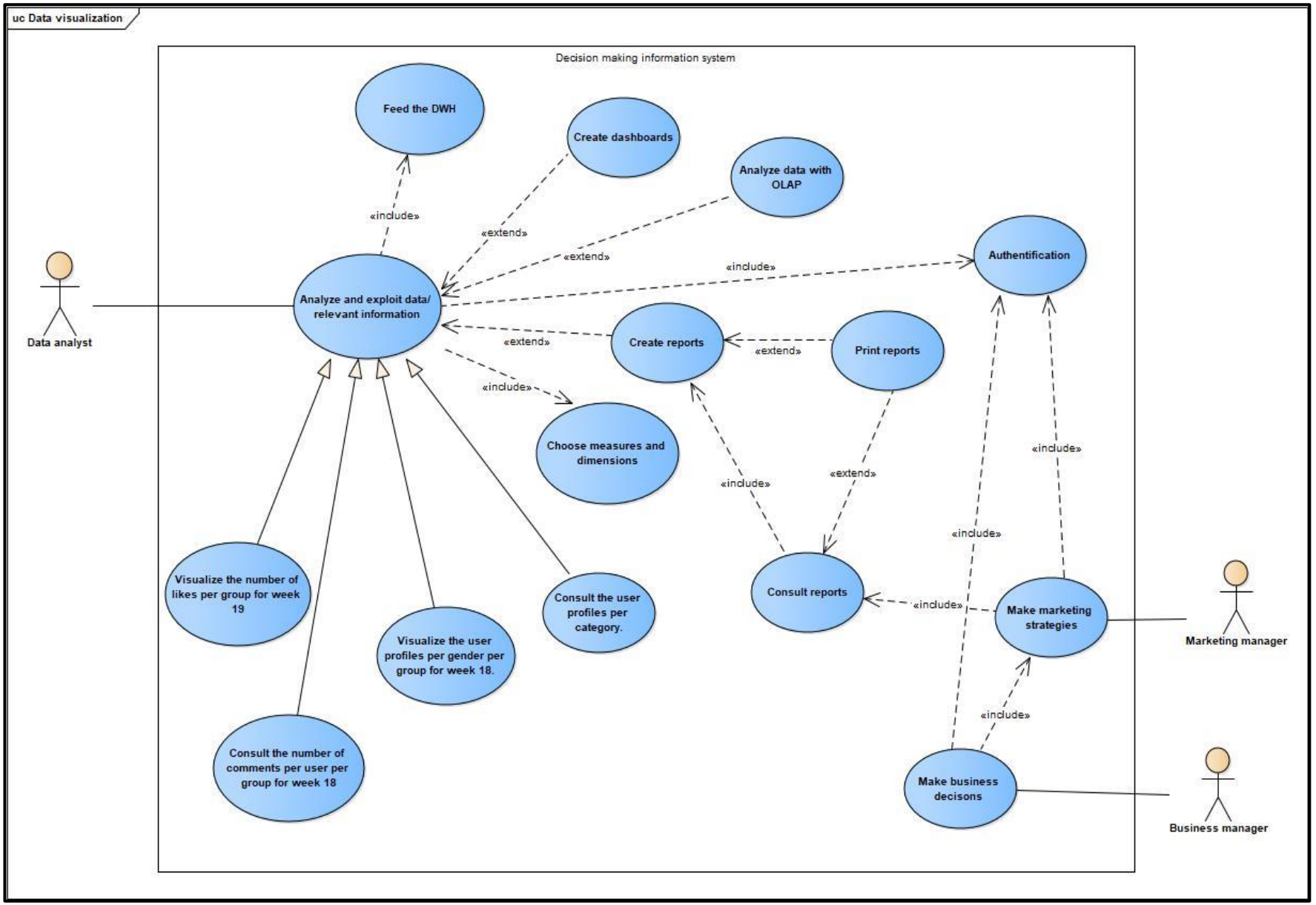
6.3. Requirement analysis

During this step, we identified the sprint's actor and described the various features/user stories with their use cases.

6.3.1. Use case diagram

In the following, we present the use case diagram "sprint1- data visualization" along with a textual description of each case/functionality.

Figure 6.1 Use case diagram-Sprint 1



6.3.2. Textual description of use cases

Table 6.2 Textual description of the use case "Analyze and exploit data/relevant information"

Title	**Analyze and exploit data/relevant information**
Resume	The system allows the data analyst to analyze the data according to specified requirements.
Actor	Data analyst
Precondition	The data analyst has the right to access the dashboard. The data analyst has a data warehouse filled with data.
Nominal Scenario	The data analyst authentificates. The data analyst selects a measure and one or more dimensions as needed to display the desired data. The data analyst launches the query. The system returns the result query.
Post-condition	Visualize the wanted results.
Complements	The data analyst must know what available data he has to work with.

Table 6.3 Textual description of the use case "Consult reports"

Title	**Consult reports**
Resume	The system allows the data analyst, the marketing manager, and the business manager to consult the reports.
Actor	Data analyst, Marketing manager, Business manager
Precondition	The actor has the right to access the dashboard.
Nominal Scenario	The actor authentificates. The actor selects the desired report to view. The system returns the report results.
Post-condition	Visualize reports.
Complements	The actor must know his needs.

6.4. Conception

The second activity in a sprint is conceptual modeling, which is translated by the sequence diagram and the class diagram. We continue with the sequence diagram to represent the interactions between the actors and the system in chronological order in this work, the same as in the previous sprints.

6.4.1. Sequence diagram

The system request to visualize data from Power BI. Then, Power BI ask SSMS to get data. SSMS send the data to Power BI and answer the system with reports.

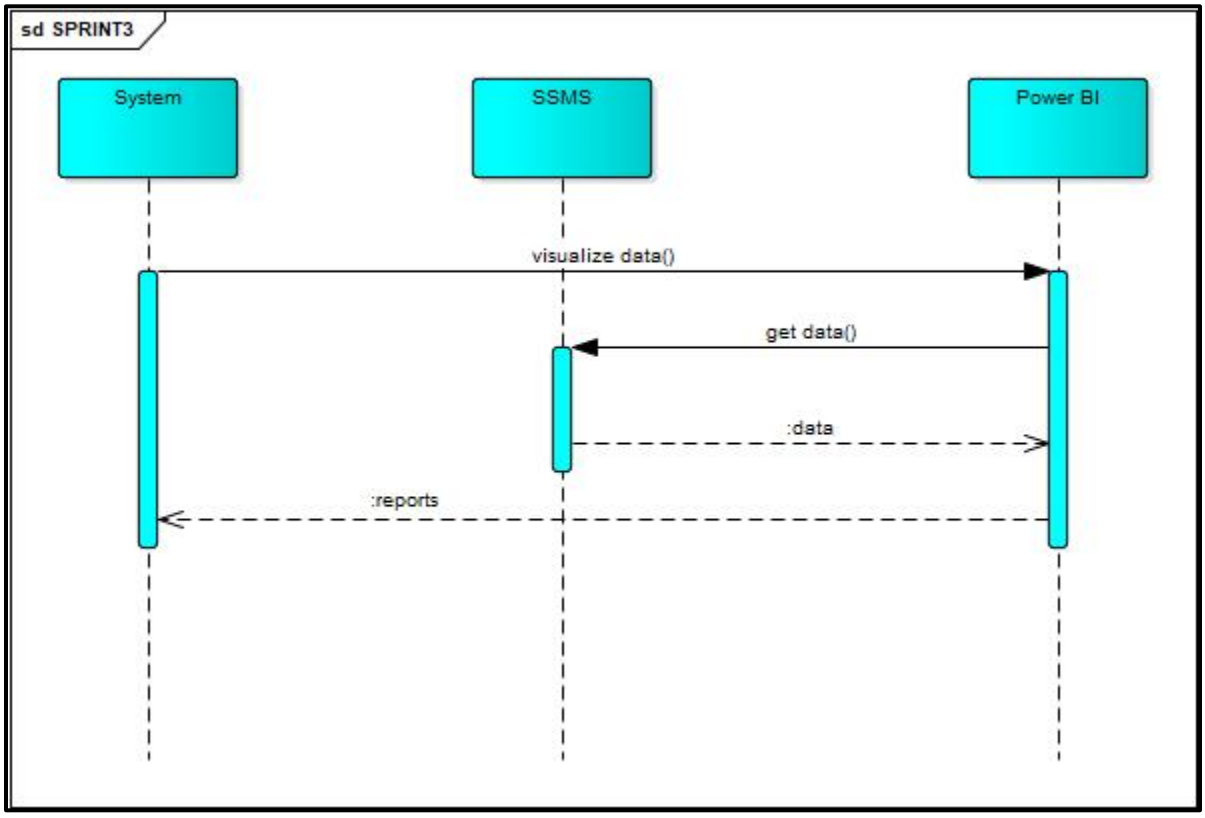


Figure 6.2 Sequence diagram- Sprint 3

6.5. Realization

In order to visualize the data, which is the objective of the project, we used the "Power BI Desktop" application that provides a dashboard representing the analysis need performed.

The dashboard is intended for the data analyst, includes the analysis of profile users and their interaction by likes and comments. Its consultation is very easy because it is equipped with an interactive interface, fast and ergonomic.

The following figure presents the interaction of likes and comments dashboard.

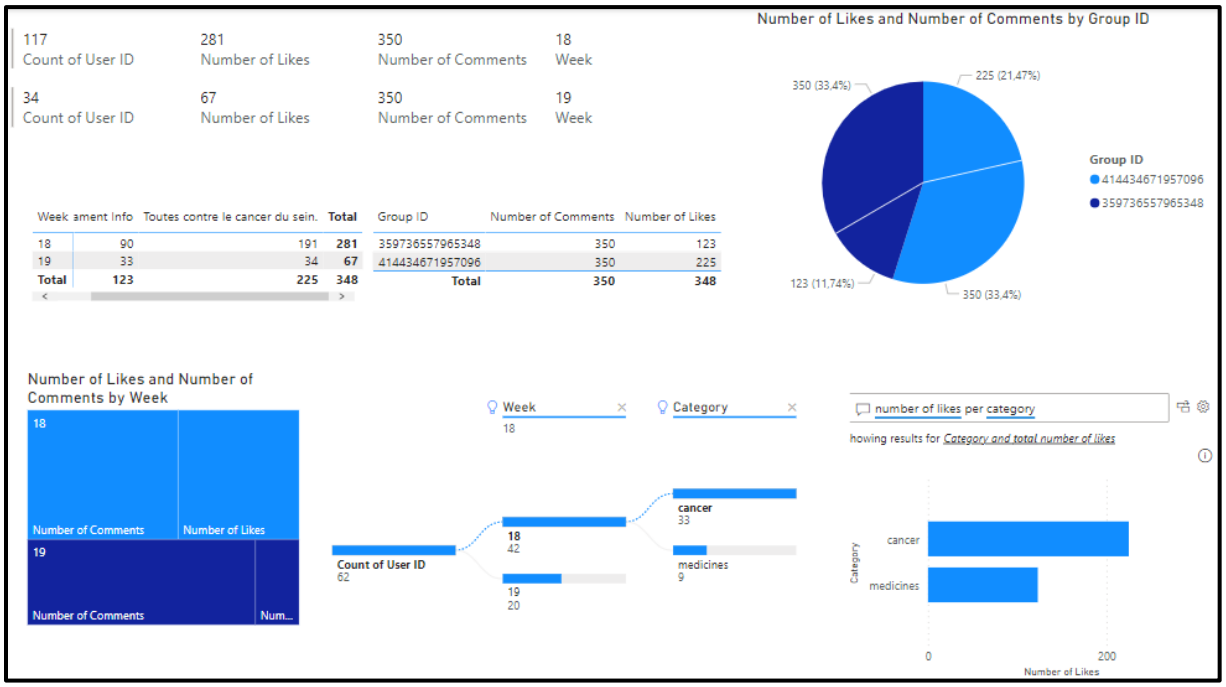


Figure 6.3 Dashboard "interaction"

We detail the dashboard related to the analysis of likes and comments:

- The multi-row card presents the number of users, the number of likes, and the number of comments to a certain week.
- The matrix shows the number of users interacting in a certain group in a certain week.
- The table demonstrates the number of comments and number of likes per group ID.
- The tree map illustrates the number of likes and number of comments per week.
- The decomposition tree presents the highest number of users per week, and per category.
- The Q&A shows the number of likes per category.
- The pie chart illustrates the number of likes and number of comments per group ID.

The figure below illustrates the user profile reports dashboard.

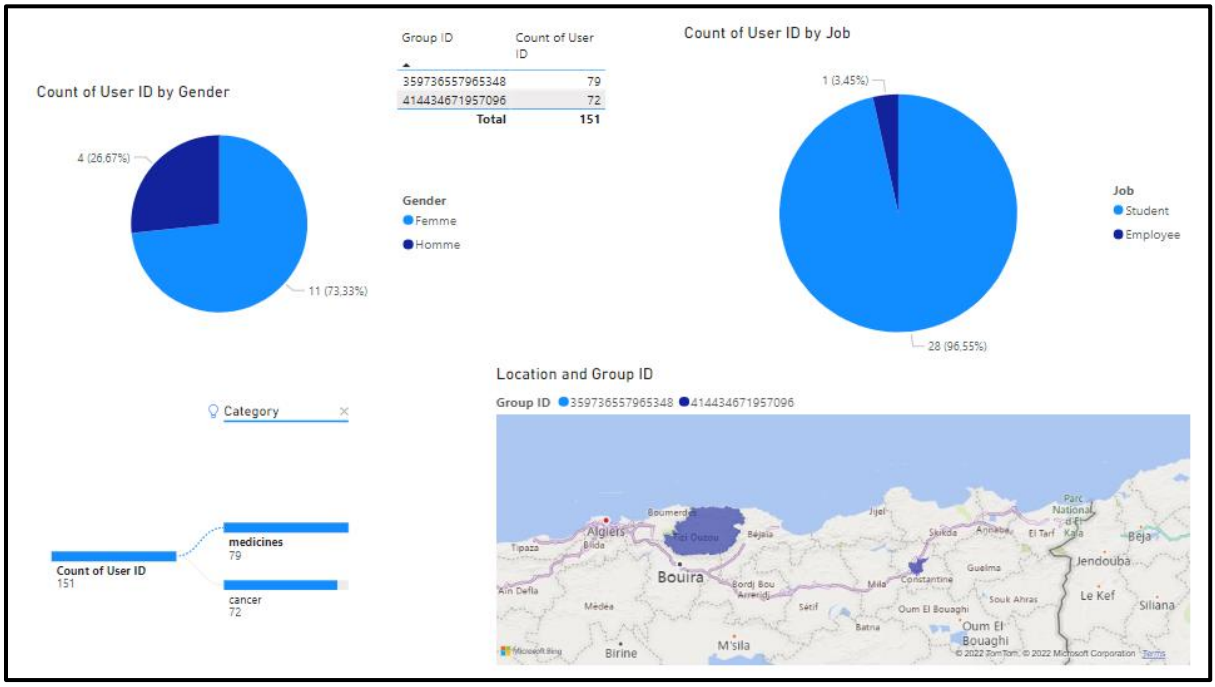


Figure 6.4 Dashboard "user profile"

We detail the dashboard related to the analysis of likes and comments:

- The donut chart illustrates the user ID by gender.
- The table presents the number of users by group ID.
- The donut chart illustrates the user ID per job.
- The decomposition tree shows the number of users per category.
- The filled map presents the location of users by groupID.

6.6. Sprint review

We present the current sprint's product increments to the scrum master Mrs. Amel BOURASSI and the product owner Mrs. Rodile ANNABI during this meeting, which is part of the sprint demonstration and validation process. We approved and validated the features of this first sprint.

6.7. Conclusion

In this chapter, we have shown the backlog of the sprint "data visualization" which describes the processed tasks.

General Conclusion

In the context of this memory, we have set the objective of developing and implementing a multidimensional schema from heterogeneous data from social networks. To achieve our goal, we developed a method for creating a data warehouse from a social network. We chose the social network Facebook because of its popularity and increasing user base. The proposed approach is very useful in that it is interesting in analyzing user interests and interactions to aid in decision-making.

According to our findings, the realization of a decision-making information system necessitates the use of a data warehouse and decision-making tools. The decision-maker can work in an informational, referenced, homogeneous, and historical environment thanks to the data warehouse. In order to construct such a system, we used Ralph Kimball's dimensional life cycle approach. The consolidated approach for building a data warehouse from Facebook groups consists of four steps. First, we collected data from Facebook group members using Phantom Buster. Then, we performed a cleaning step to determine the multidimensional concepts to generate the data warehouse schema. Then, we defined the loading procedures. Finally, we realized the visualization of the data. We used the agile framework "Scrum" to develop our project, which allowed us to work in an iterative and incremental manner.

It is necessary to mention at the end of our study our delight at having acquired so much knowledge in the BI and Big Data domains. This project provided an excellent opportunity to perform concrete work while also becoming acquainted with the workplace and professional environment. We were able to put our theoretical knowledge into practice, but we were also able to improve, perfect, and adapt our knowledge to the needs of the companies by overcoming the limitations discovered.

The technical aspects of our work are not completely perfect and could be improved eventually. In terms of perspectives, we suggest, on the one hand, to evaluate our design process on other social on other social networks, on the other hand, to extend the operations proposed in the ETL process by operations specific to the context of social networks.

REFERENCES

[1]: Agile model: https://a-connect.com/knowledge/agile-the-art-of-adapting-to-change-and-innovation/

[2]: Agile model: 15th Annual State of Agile Report, 2021

[3]: Scrum: Ken Schwaber & Jeff Sutherland, 2020, *The Scrum Guide The Definitive Guide to Scrum: The Rules of the Game*.

[4]: Business Intelligence: https://www.supinfo.com/articles/single/3548-comprendre-etapes-processus-bi

[5]: Business Intelligence benefits: https://bi-survey.com/benefits-business-intelligence

[6]: Social media: K.Tietze & T.Schlegel, 2011, "on modeling a social networking service description", *Gemeinschaften in Neuen Medien: Virtual Enterprises, Communities & Social Networks Workshop (GeNeMe '11), TUDresden Edition*

[7]: Social media: D.M.boyd & N.B.Ellison, 2007, "Social Network Sites: Definition, History, and Scholarship", *computer-mediated communication journal, Published by the international communication Association (IA'07),* vol 13 No.1, pp 210-230.

[8]: Facebook: https://en.wikipedia.org/wiki/Facebook

[9]: Twitter: https://en.wikipedia.org/wiki/Twitter

[10]: Data warehouse: https://www.lebigdata.fr/data-warehouse-entrepot-donnees-definition

[11]: Data Dimensional Modeling: https://www.guru99.com/dimensional-model-data-warehouse.html

[12]: Data marts: https://en.wikipedia.org/wiki/Data_mart

[13]: On-Line Analytical Processing: https://en.wikipedia.org/wiki/Online_analytical_processing

[14]: DWH modeling approaches : Ghozzi Faiza, 2004, "CONCEPTION ET MANIPULATION DE BASES DE DONNEES DIMENSIONNELLES À CONTRAINTES", *Informatique [cs]*.

[15]: OSINT: https://en.wikipedia.org/wiki/Open-source_intelligence

[16]: User profile: Marina Farid, Rania Elgohary, Ibrahim Moawad & Mohamed Roushdy, 2018, "User Profiling Approaches, Modeling, and Personalization", *Proceedings of the 11th International Conference on Informatics & Systems (INFOS 2018).*

[17]: UML: https://en.wikipedia.org/wiki/Unified_Modeling_Language

[18]: Phantom Buster: https://www.getapp.com/business-intelligence-analytics-software/a/phantombuster/

[19]: VS code: https://code.visualstudio.com/docs/supporting/FAQ

[20]: PostgreSQL: https://www.postgresql.org

[21]: Talend: https://fr.wikipedia.org/wiki/Talend

[22]: SSMS: https://www.techtarget.com/searchdatamanagement/definition/Microsoft-SQL-Server-Management-Studio-SSMS

[23]: Power BI Desktop: https://powerbi.microsoft.com/en-us/desktop/

[24]: Enterprise architect: https://www.techopedia.com/definition/7046/enterprise-software-architecture

[25]: Python: https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/

[26]: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[27]: https://www.statista.com/statistics/1176789/uk-most-important-mobile-apps/

[28]: https://www.statista.com/statistics/1015131/impact-of-social-media-on-daily-life-worldwide/

[29]: https://gs.statcounter.com/social-media-stats/all/africa