A black and white logo

Description automatically generated

Data Catalog as a SaaS Solution

Report

2023 – 2024

Supervisors: Elaborated by:

Rihab Idoudi Yacine Tazarki

Thouray Louati Haithem Meddeb

Med Ali Belhadj

Rahma Farhat

Med Houssem Rejeb

Contents

[Chapiter 0 : Sprint 0 3](#_Toc160729225)

[I. Introduction 3](#_Toc160729226)

[II. Preliminary Study 4](#_Toc160729227)

[1. Goals 4](#_Toc160729228)

[2. State of the Art 4](#_Toc160729229)

[3. Proposed Solution 4](#_Toc160729230)

[4. Work Plan 4](#_Toc160729231)

[III. Planning the Application 5](#_Toc160729232)

[1. Requirement Analysis 5](#_Toc160729233)

[Functional Requirements 5](#_Toc160729234)

[Non-Functional Requirements 5](#_Toc160729235)

[2. Technical Requirements 9](#_Toc160729236)

[Technical Architecture 9](#_Toc160729237)

[Languages and Frameworks 11](#_Toc160729238)

[3. Prototyping 11](#_Toc160729239)

[Logo 11](#_Toc160729240)

[User Interface 11](#_Toc160729241)

[VI. Conclusion 14](#_Toc160729242)

[Chapiter 1 : Sprint 1 15](#_Toc160729243)

[I. Introduction : 15](#_Toc160729244)

[II. Columns module : 15](#_Toc160729245)

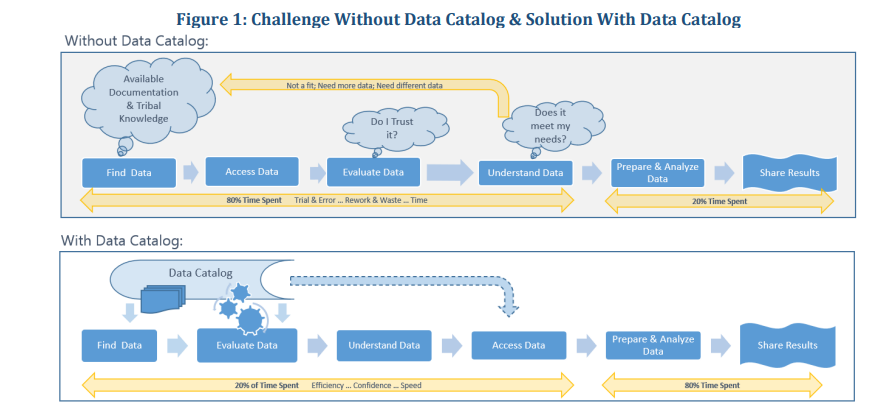
[1. Sprint 1 backlog : 15](#_Toc160729246)

[2. Use Case Diagram : 17](#_Toc160729247)

# Chapiter 0 : Sprint 0

## I. Introduction

In the rapidly evolving landscape of data analytics and enterprise solutions, the need for effective data governance has become paramount. This report outlines the development plan for a Minimum Viable Product (MVP) of a Data Catalog SaaS application, catering to the data governance needs of a French client. The application aims to address challenges related to data control, responsibility, traceability, and communication within organizations.



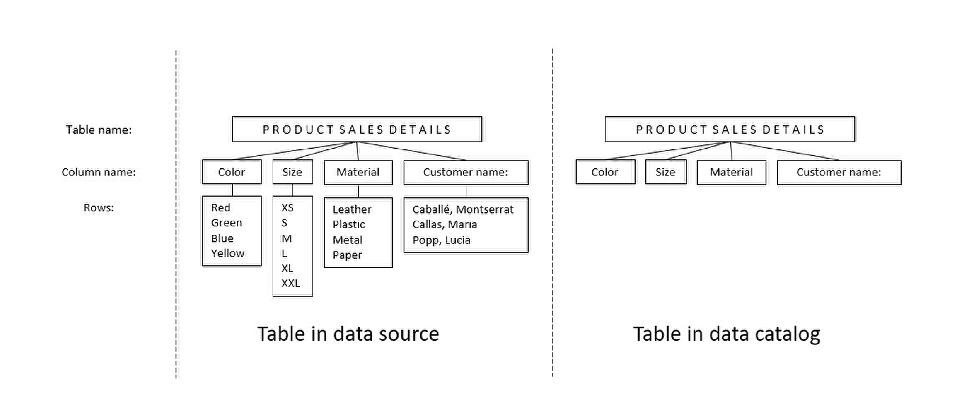


Figure 2: Extracting metadata from data

## II. Preliminary Study

## 1. Goals

The primary goal of the project is to develop a comprehensive Data Catalog SaaS application that serves as the core of data governance. The application should allow for the manual and automatic cataloging of metadata, enabling efficient data management, categorization, and access control. Additionally, it should act as an innovative solution for data analytics.

## 2. State of the Art

A thorough analysis of existing tools and platforms for data governance has been conducted. This includes an evaluation of features, functionalities, and usability to identify gaps in current solutions that the proposed application can address. The goal is to leverage the strengths of existing tools while introducing innovative features to meet the unique requirements of the client.

## 3. Proposed Solution

The proposed solution is the development of a Data Catalog SaaS application that centralizes, categorizes, and traces metadata. It offers customization capabilities, automates data analysis and entry, provides data traceability visualization, integrates with external systems through APIs, and stores data quality/profiling information.

## 4. Work Plan

The work plan involves several phases:

1. **Research and Evaluation:**
   * Evaluate existing data governance tools.
   * Understand the landscape to identify opportunities for innovation.
2. **Architecture Design:**
   * Design a scalable and secure architecture using Spring Boot for the backend, Angular for the frontend, and Docker for containerization.
3. **MVP Development:**
   * Develop the MVP with a focus on centralizing and entering data, customization, and API integration.
4. **Machine Learning Module:**
   * Develop a machine learning model for workload projection and data quality monitoring.
5. **Testing and Iteration:**
   * Conduct thorough testing.
   * Iterate based on feedback and testing results.
6. **Documentation:**
   * Create comprehensive documentation for users and developers.
7. **Future Considerations:**
   * Plan for future development, scaling, and integration with larger projects.

## III. Planning the Application

## 1. Requirement Analysis

### Functional Requirements

The functional requirements include:

* Centralizing and entering data.
* Customization of software and data information.
* Automation of data analysis and entry.
* Visualization of data traceability and access.
* Integration through API.
* Storage and visualization of data quality/profiling.

### Non-Functional Requirements

1. **Security:**

Robust security measures will be implemented to safeguard sensitive data. This includes data encryption, secure API communication, and access controls.

1. **Performance:**

The application must deliver responsive performance, ensuring quick data access and analysis. Response times for critical functionalities should meet defined benchmarks.

1. **Reliability:**

The system will exhibit high reliability, minimizing downtime, and ensuring consistent availability to users.

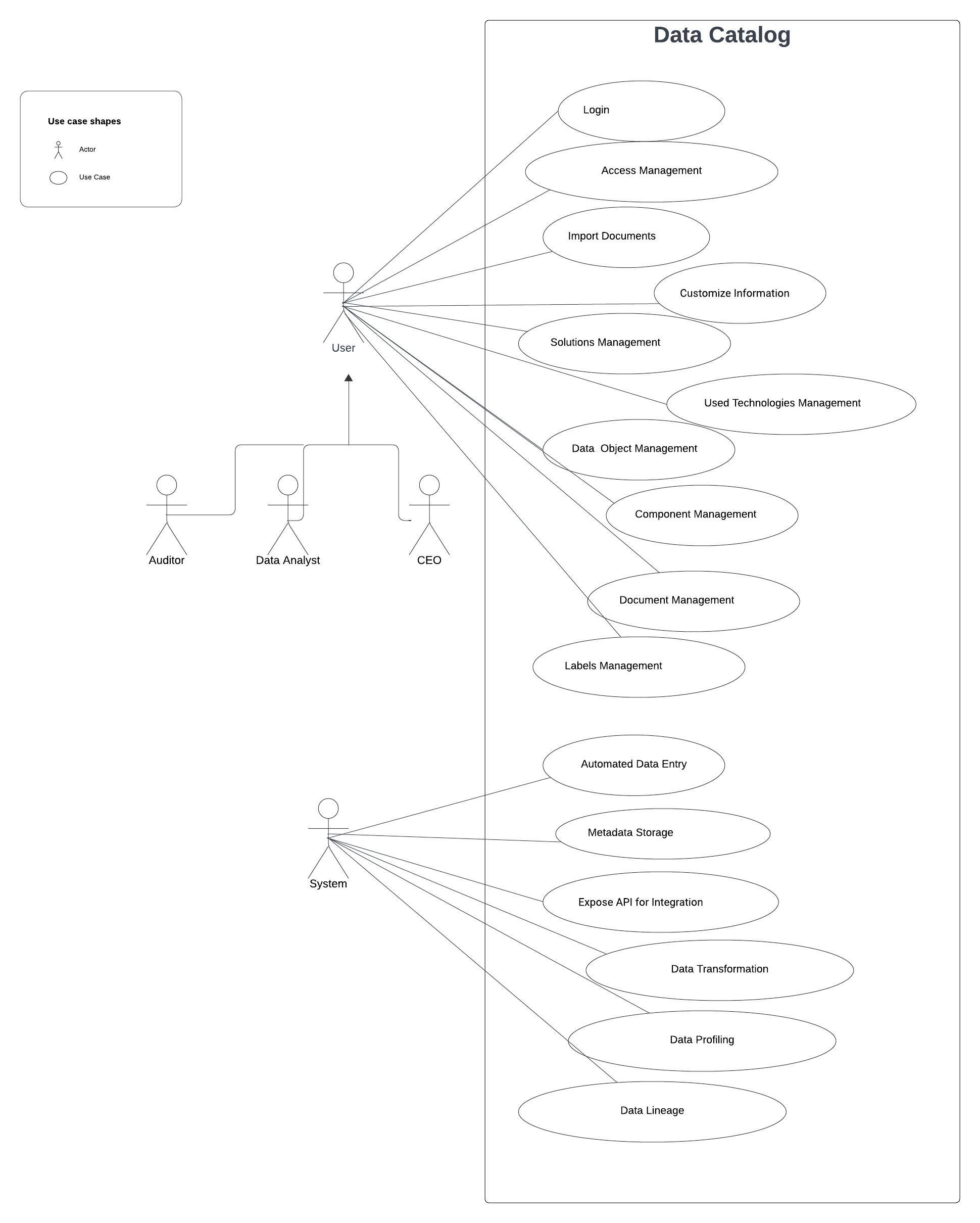


Figure 3: Use Case Diagram

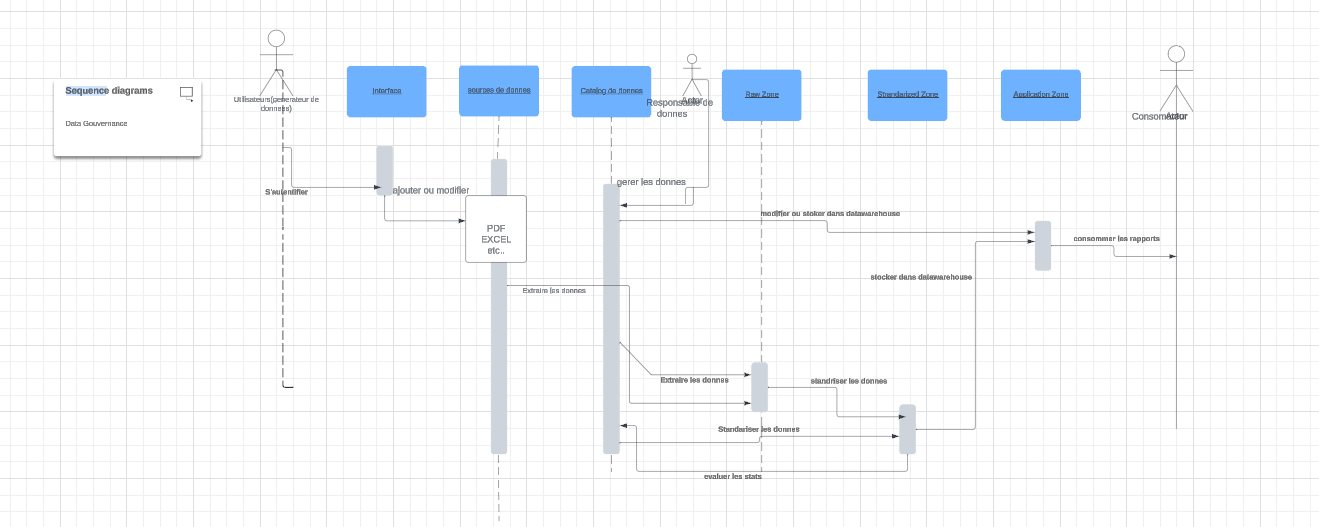
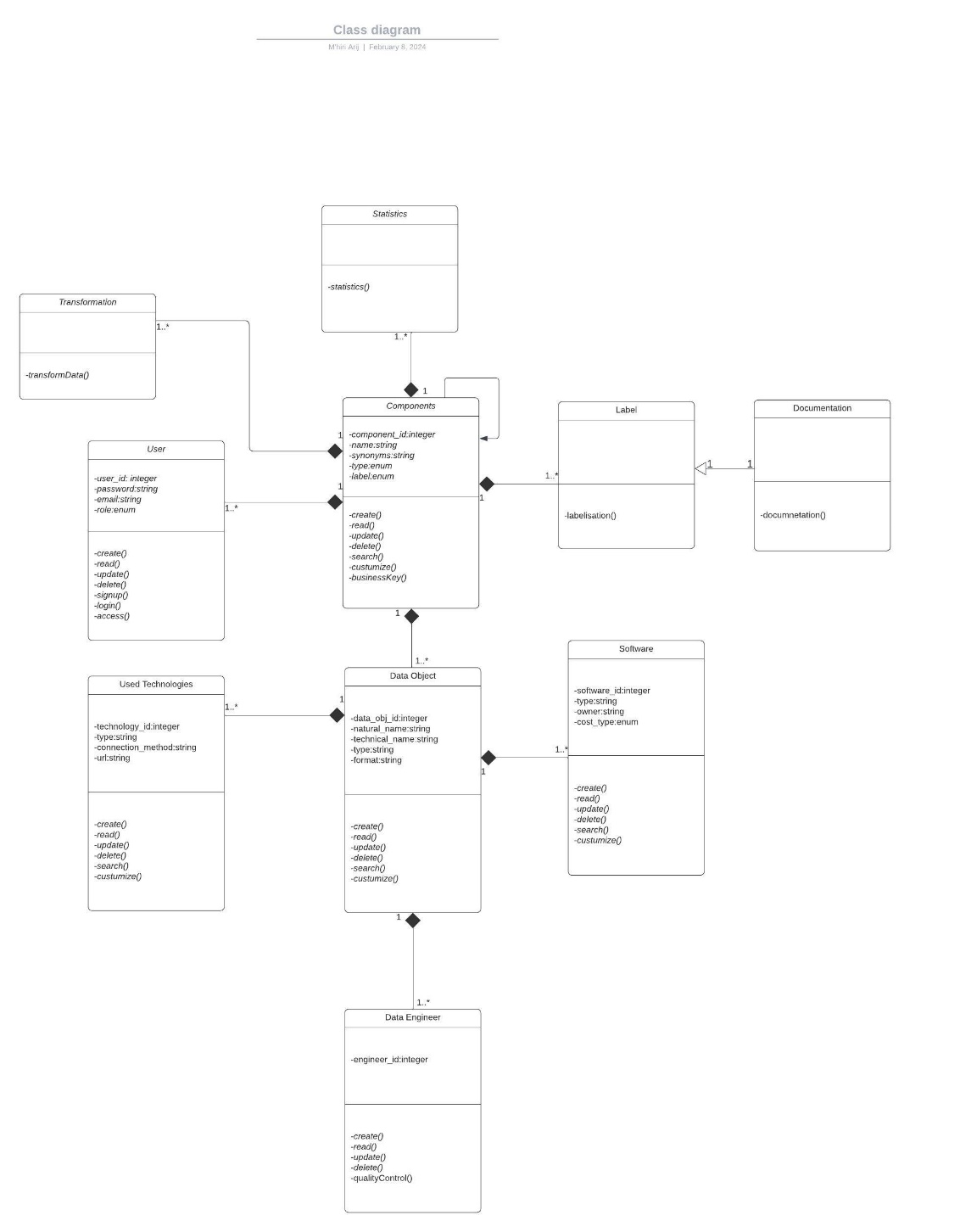


Figure 5: Class Diagram

Figure 4: Global Sequence Diagram

Figure 6: Class Diagram

## 2. Technical Requirements

### Technical Architecture

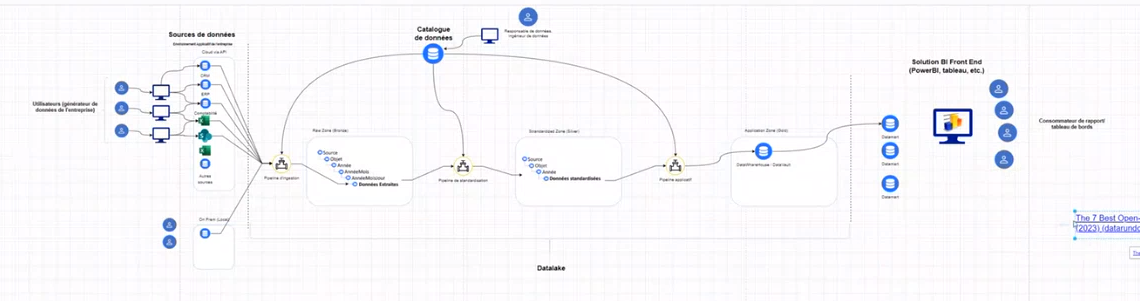
The application will be built on a microservices architecture with a Spring Boot backend serving as the API gateway. Angular will be used for the frontend, providing a responsive and user-friendly interface. Docker containers will be employed for deployment, ensuring consistency across different environments.

Figure 7: Company Setting

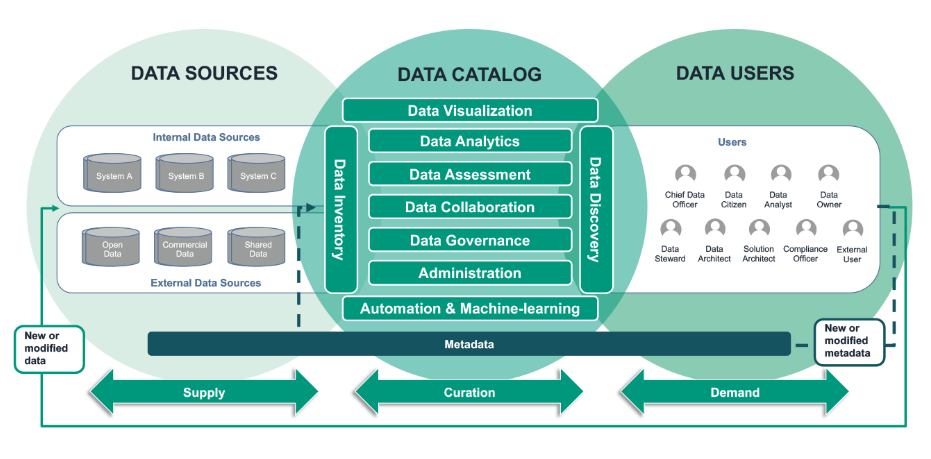


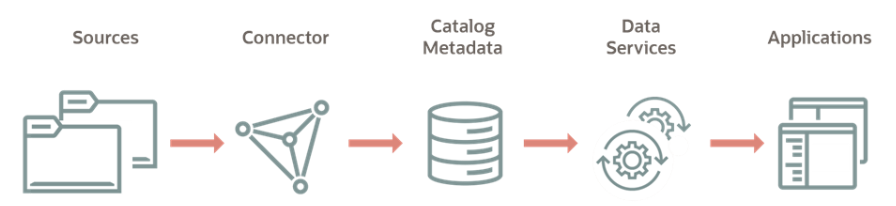
Figure 8: Data Catalog Architecture

#### A diagram of a presentation Description automatically generatedLogical Architecture

A diagram of a model

Description automatically generated

#### Technical Architecture



### Languages and Frameworks

* **Backend:**

Java with Spring Boot for robust and scalable API development.

Integration with MongoDB and MySQL for data storage.

* **Frontend:**

Angular for a dynamic and interactive user interface.

* **Containerization:**

Docker for containerization to enhance portability and deployment efficiency.

* **Integration:**

Utilization of APIs for seamless integration with external systems.

* **Machine Learning:**

Selection of appropriate machine learning libraries for workload projection and quality monitoring.

By adhering to these technical requirements, the application aims to achieve a balance between innovation, scalability, and user experience.

## 3. Prototyping

### Logo

Figure 9: Logo

### User Interface

A computer with a login screen

Description automatically generated

Figure 10: Login page

A computer screen with text on it

Description automatically generated

Figure 11: Table of users



Figure 12: Example of categories table

A computer with graphs on it

Description automatically generated

Figure 13: KPI Dashboard

A computer with a chat window

Description automatically generated

Figure 14: Data lineage

## VI. Conclusion

In conclusion, this report outlines the goals, state of the art, proposed solution, and work plan for the development of a Data Catalog SaaS application. The technical requirements provide a roadmap for creating a robust and user-friendly application that addresses the challenges of data governance. The subsequent phases will focus on detailed planning, development, and testing to deliver a high-quality MVP that aligns with the client's objectives.

# Chapiter 1 : Sprint 1

### Introduction :

The product will be developed using Sprints, which have a consistent duration and never overlap. This means that one Sprint cannot begin until the previous one has been completed. Before starting a Sprint, the Scrum team must mandatorily define the Sprint's goal. This goal should be defined in business terms, not technical terms, to ensure that it is understandable by members outside the team. It is essential to answer a fundamental question: "Why are we doing this Sprint?" Following a conversation between the Product Owner and the Scrum team, the following goal has been decided: "To complete the part related to project management, modules, risks, and budgets after the creation of the application's architecture and database as well as the design."

## Columns module :

### 1. Sprint 1 backlog :

A backlog is a prioritized list of tasks, issues, or user stories that need to be addressed in a software development project, specifically tailored to the provided context. In this case, the backlog represents a collection of tasks or issues within the columns module :

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### Use Case Diagram :

The context describes the roles, responsibilities, and interactions within a data system. The primary roles in the system include : Data Protection Officer, Data Analyst, and CEO.

The Data Protection Officer and Data Analyst, as well as the CEO, are responsible for reading the documentation provided for the system. The Data Protection Officer, and possibly the CEO, have the additional capability to modify the documentation.

The Column Module allows users to manage columns, including adding new columns, reading all columns, updating a column, deleting a column, and searching the column table by label. This Column Module enables users to effectively interact with the data and tailor the table to their needs while maintaining a secure and organized data structure.

