Tunisian Republic Ministry of Higher Education And Scientific Research

University of Sfax

The National School of Electronics and Telecommunications of Sfax

Engineer in :
Data Engineering and Decisional System

Option : Data Science
End of Studies Project
Order N° : IDSD-SD-3-2025-006

OIP -FR-01-00
05/05/2025

Graduation Report

Presented at

The National School of Electronics and Telecommunications of Sfax

In order to obtain the

Engineering Diploma in:
Data Engineering and Decisional System

Option:

Data Science

By

Ghaith KRIFA

Tenders Observatory: News & Tenders Opportunities

Presented on 18/06/2025, before the review panel

Mrs.	**Ines BEN MASSOUD**	President
Mr.	**Mohamed KOUBAA**	Examiner
Mr.	**Tarak BEN SAID**	Academic Supervisor
Mr.	**Farouk JAZIRI**	Industrial Supervisor

DEDICATION

To my dear and beloved parents Brahim and Feiza, I would like to extend my heartfelt dedication, as no words can truly express the depth of my love, immense gratitude, and utmost respect for them. From my earliest years, they have enveloped me with their unwavering tenderness and unwavering love.

I would also like to dedicate this humble report to my entire family,

including my brothers Ahmed, Anas and Hamza

for their support and encouragement.

To all my friends, loved ones, and all those who have provided me with encouragement and support during this internship. I am truly grateful for their presence in my life.

To my esteemed teachers, who have guided and nurtured me throughout my academic journey.

To all the members of the jury who have invested their time

and expertise in evaluating my work.

To the entire staff of "DNEXT" who have played an instrumental role in helping me achieve my goals,

to all the students of ENET'COM.

And to you dear readers,

Thank you all!

Ghaith KRIFA

ACKNOWLEDGEMENTS

I would like to begin by expressing my deepest gratitude to my friends and colleagues for their steadfast moral and intellectual support throughout this project. Your encouragement has been truly invaluable.

I am also profoundly thankful to the entire team at “Talan Tunisie” for their assistance, guidance, and collaboration. Your collective efforts have been instrumental in the success of this work.

My sincere appreciation goes to **Mr. Farouk JAZIRI**, my supervisor at "DNEXT" for his insightful guidance and constant motivation. Your leadership and generosity inspired me to give my best. I would also like to extend my gratitude to **Mr. Jebrane BEN SALEM**, our Head of Data Analysis Team in Tunisia, for his valuable support and expertise throughout this journey.

A special thank to **Mr. Tarek BEN SAID**, my academic supervisor, for his exceptional teaching, encouragement, and unwavering support during every stage of this project. I am deeply grateful for the time and dedication you invested in me.

I would also like to thank the members of the jury for the honor of reviewing and evaluating this work. Your time and expertise are greatly appreciated.

Furthermore, I wish to acknowledge the pedagogical and administrative staff of ENET’COM for providing an excellent academic foundation and fostering a supportive learning environment.

Finally, I extend my thanks to everyone who contributed, in big ways and small, to the realization of this project. Your support has made a meaningful difference, and I am truly thankful.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

API	Application Programming Interface
CNF	Cost and Freight
DAG	Directed Acyclic Graph
FOB	Free On Board
KPI	Key Performance Indicator
LLM	Large Language Model

GENERAL INTRODUCTION

In today's fast-paced global economy, tenders and market opportunities are crucial for shaping strategic trade decisions across various industries. Access to accurate and timely tender data enables businesses to make informed decisions that drive economic growth, optimize supply chains, and maintain competitiveness. Stakeholders who can quickly adapt to market changes can seize new opportunities, mitigate risks, and enhance operational efficiency. Delays or inaccuracies in managing tender data can lead to missed opportunities, inefficiencies, and decreased responsiveness, ultimately hindering economic progress.

In the agricultural sector, especially within the grains and oilseeds market, timely and reliable tender data is vital for guiding trade and production decisions. Tenders in this area directly influence the pricing, sourcing, and distribution of these key agricultural products, impacting global trade flows and market supply chains. With up-to-date tenders information, agricultural businesses can make smarter decisions to optimize operations, manage risks, and respond to changes in market demand. By staying ahead of price fluctuations and supply shifts, stakeholders in grains and oilseeds can maintain a competitive advantage in an increasingly globalized market.

This project is part of the requirements for obtaining an engineering diploma in Data Engineering and Decision Systems from the National School of Electronics and Telecommunications of Sfax. Conducted as an internship at DNEXT, a leading international company specializing in agricultural commodity market intelligence, this project addresses the challenge of manual data collection that slows down analysis updates. DNEXT required a scalable solution capable of processing large volumes of tender information quickly while maintaining data consistency and accuracy. The system encompasses end-to-end automation: extraction,quality control, and visualization.

This report is organized into five main chapters:

- The first chapter provides an overview of the project's general context, including the business background, and the methodology for project implementation during the internship.

- The second chapter outlines the preparatory phase of the project, beginning with an analysis of the current context, identified limitations, and the proposed solution. It also covers the technical and functional requirements, the SCRUM-based project management approach, and the description of the work environment.

- The third chapter presents an overview of the developed workflows for each data source, detailing the AI models integrated within them and their performance evaluation.

- The fourth chapter explains the verification of data quality before storing it in the database, describing the workflow steps and the involvement of senior experts to ensure data accuracy.

- The fifth chapter covers the deployment of workflows using Apache Airflow and Docker, as well as the creation of interactive Power BI dashboards for stakeholder insights from processed tenders.

We hope that this project offers valuable insights and decision-aid to enhance the understanding and strategic decision-making processes.

Chapter

1

Project General Context

Contents

1.1 Introduction

This chapter offers a comprehensive overview of the overall context of my internship project. It begins with a presentation of the host company, outlining its missions and market focus. Then, it delves into the business context, introducing key concepts essential to understanding the project's scope. The chapter concludes by detailing the methodology adopted to guide the implementation of the project.

1.2 Hosting Company

This section provides an overview of DNEXT, focusing on its role in the agricultural commodities sector and its approach to delivering market intelligence insights.

1.2.1 Company Presentation

Founded in 2020, DNEXT is a dynamic Switzerland-based firm specializing in agricultural commodity expertise and market intelligence solutions. In 2023, it generated around 5 M€ in annual revenue and employed more than 50 people. In a short time, the company has emerged as an influential player in the global agricultural commodities sector. It provides comprehensive insights that enable clients to identify market opportunities, understand evolving trends, enhance risk-management strategies, and optimize decision-making.

Figure 1.1: Company Logo

1.2.2 Company Missions

DNEXT has a global presence with offices in Switzerland, the United Kingdom, Spain, Tunisia, India, and Brazil, ensuring proximity to key agricultural production regions and markets worldwide. This international presence allows DNEXT to provide localized market insights and responsive support tailored to regional market dynamics and client-specific requirements. The company is deeply committed to innovation and continually enhances its capabilities to provide strategic insights and reliable market intelligence. DNEXT's proactive approach empowers clients with informed decision-making capabilities, driving sustained operational success and ensuring a competitive advantage in the global agricultural marketplace.

1.2.3 Market Focus

DNEXT focuses on various aspects of agricultural market intelligence. Below are the key areas of its analytical focus:

- **Weather:** Incorporates weather data into its analysis to predict crop yields and assess market conditions in agriculture.
- **Tradeflow:** Analyzes global trade flows of agricultural commodities, including imports, exports, and trade dynamics.
- **Grains & Oilseeds:** Specializes in insights into global grains and oilseeds markets, covering pricing, supply, and demand trends.
- **Sugar:** Provides market analysis and forecasts for the sugar industry, an essential agricultural commodity.
- **Biofuels:** Examines biofuel production and its impact on agricultural markets, particularly regarding feedstock crops such as corn, sugarcane and soybeans.
- **Fertilizers:** Analyzes fertilizer markets to understand how input costs and availability affect agricultural productivity.
- **Meat:** Covers livestock and meat markets, which are closely tied to grain and oilseed demand due to animal feed requirements.

1.3 Business Understanding

To fully understand the project, there are several key aspects that need to be addressed. In this section, we aim to explain the most important ones.

1.3.1 Tender

When a company is interested in purchasing something via tenders, it follows these steps:

- **Announcement:** The company publishes details about what it intends to purchase, including specifications.
- **Offers:** Suppliers submit their bids based on the company's specifications.
- **Results:** The company evaluates the offers and announces the selected ones.

In our business, we specialize in tracking tenders related to grains and oilseeds. with a strong focus on analyzing tender results and offers, as announced tenders can sometimes be canceled or modified.

1.3.2 Freight Charge

Refers to the cost of transporting goods from one location to another, typically via a shipping method such as sea, air, road, or rail. These charges can vary depending on several factors, including the distance, weight, volume of goods, transportation mode, and the service provider. Freight charges typically cover the costs of moving goods between the seller's and buyer's locations.

1.3.3 Price Types

- **Basis Price:** The basis price refers to the starting price for a product or service, but it can change depending on certain factors like market conditions or material costs.

- **Flat Price:** A flat price is a fixed price that does not change. No matter what happens, you pay the same price for the product or service.

1.3.4 Delivery Terms

- **CNF:** The seller is responsible for the cost of transportation to the destination port, but the risk transfers at the loading point. [2]

- **FOB:** The seller's responsibility ends once the goods are on board the vessel, with the buyer assuming both risk and cost of the goods from that point onward. [2]

Figure 1.2 below explains the different delivery terms commonly used in trade agreements:

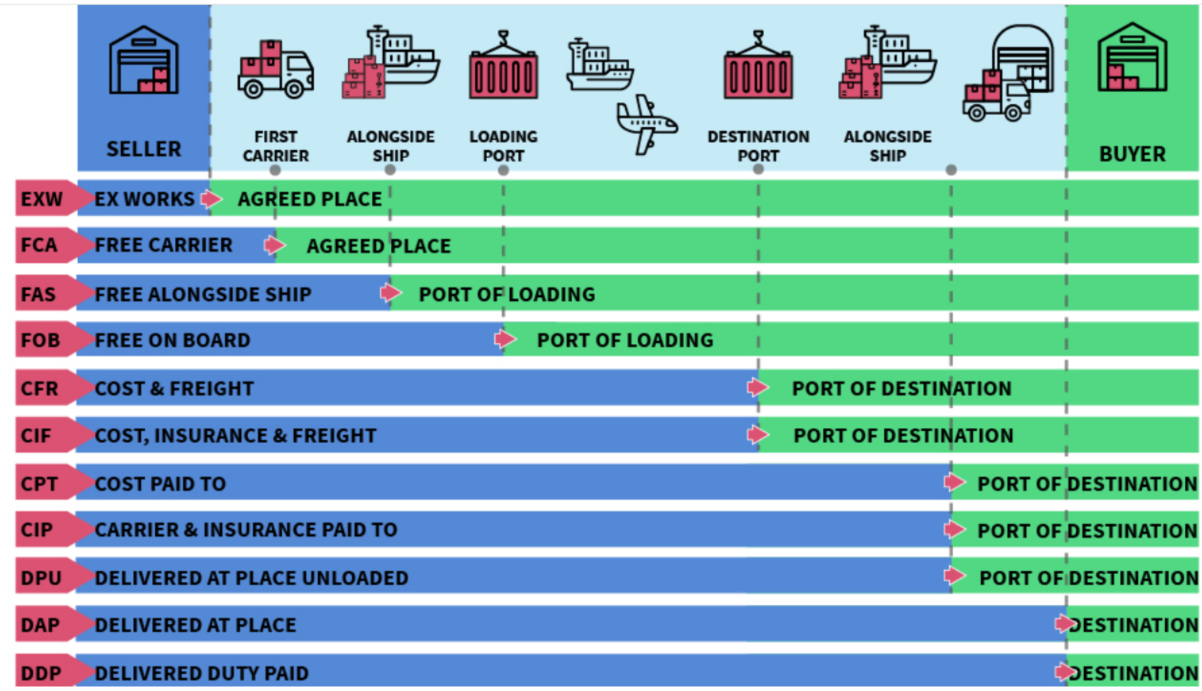


Figure 1.2: Shipping terms

1.4 Project Methodology

This section outlines the adopted methodology for the project, focusing on the Agile Scrum framework. It discusses the key steps involved in Scrum and the roles within a Scrum team, highlighting its iterative approach to project management and its emphasis on flexibility, collaboration, and continuous improvement.

Agile methodologies are characterized by their flexibility and ability to adapt to change. They provide development processes that promote collaboration and communication with the client. Scrum is one of the most widely recommended agile methodologies. It is an iterative and incremental framework that allows teams to complete a set of tasks within a defined time interval called a sprint.

1.4.1 Scrum Steps

- **Sprint Planning:** This is a meeting where the team defines the goals for the upcoming sprint.
- **Daily Stand-up:** It is a daily meeting to discuss progress and any obstacles.
- **Sprint Execution:** The team collaborates to accomplish the goals defined during the sprint planning.
- **Sprint Review:** At the end of each sprint, the team conducts a presentation to validate the increment achieved.
- **Sprint Retrospective:** It is a meeting at the end of each sprint to identify the positives and discuss potential improvements for future sprints.

1.4.2 Scrum Roles

The members of a SCRUM team are divided into three different SCRUM roles, with no hierarchy between them.

- **The SCRUM Master:**
  - Ensures that the Scrum framework is respected.
  - Commits to upholding Scrum values and practices.

  - Seeks to improve the team's productivity and expertise.
  - Facilitates communication within the team.

- **The Team:**
  - All team members contribute their expertise to complete tasks.
  - Typically consists of 3 to 9 people.
  - No fixed roles such as architect, developer, or tester.

- **The Product Owner:**
  - Business expert who defines the functional specifications.
  - Sets feature development priorities and manages the product backlog

In summary, the steps of the Scrum methodology provide a solid framework for agile, iterative, and incremental project management. We have adopted this methodology for our project. Figure 1.3 below provides an overview of the Scrum life cycle:

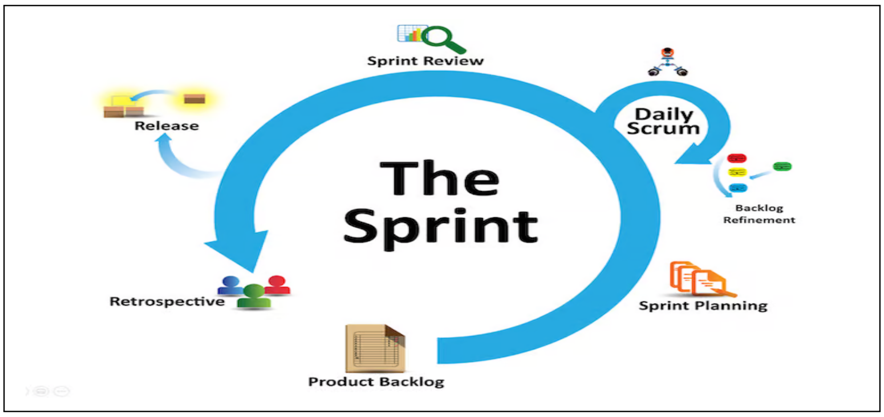


Figure 1.3: Scrum Life Cycle

1.5 Conclusion

In conclusion, this chapter has laid a solid foundation by presenting the host company and its strategic focus, as well as the broader business context relevant to the project. By clearly defining the scope and introducing the guiding methodology, it sets the stage for the successful execution and management of the internship project ahead.

Chapter

2 Project Preparation

Contents

2.1 Introduction

This chapter presents the preparatory phase of the project. It begins with an analysis of the project, outlining the current context, the limitations of the existing work, and the proposed solution to address these challenges. It then moves on to a comprehensive analysis and specification of the technical and functional requirements. The chapter also details the project management approach based on the SCRUM methodology, including the creation of the product backlog and the sprint planning process. Finally, it describes the work environment, covering both hardware components and the development tools necessary for the successful execution of the project.

2.2 Project Analysis

This section provides an overview of the current process for tender data collection, its limitations, and how automation can improve data quality, reduce errors, and enhance client value.

2.2.1 Overview of Existing Work

- **Text Analysis and Data Collection:**

  The team begins by searching for tenders across multiple sources. Most sources provide tender information in text format that requires careful analysis to extract key information. The team must analyze these text-based tender documents and compile approximately 15 key details for each tender, such as origin, product, quantity, price, dates, and delivery terms.

- **Conflict Resolution for Tender Information:**

  When conflicting details about the same tender appear across sources, the team emails the senior analyst, who decides which version to use.

- **Verification Across Data Sources:**

  After initial data extraction, the team verifies this information across all available sources to identify if there is missed information that is mentioned in other sources and fill it. Additionally, they verify that new tenders have been published in other sources that need to be processed and extracted.

2.2.2 Limitations

the collection of tender data is still mostly manual and involves several complex steps. This manual process consumes significant time, delaying real-time updates to our dashboards. Moreover, it introduces risks of human error, such as missed tenders or inaccuracies in extracted data. These issues ultimately compromise service quality by risking the delivery of outdated or incorrect insights to clients.

2.2.3 Proposed Solution

The following section outlines the key components of the proposed solution, which aims to streamline the data collection and processing workflow for enhanced efficiency and accuracy. Figure 2.1 below presents an overview of the designed workflow for the proposed solution:

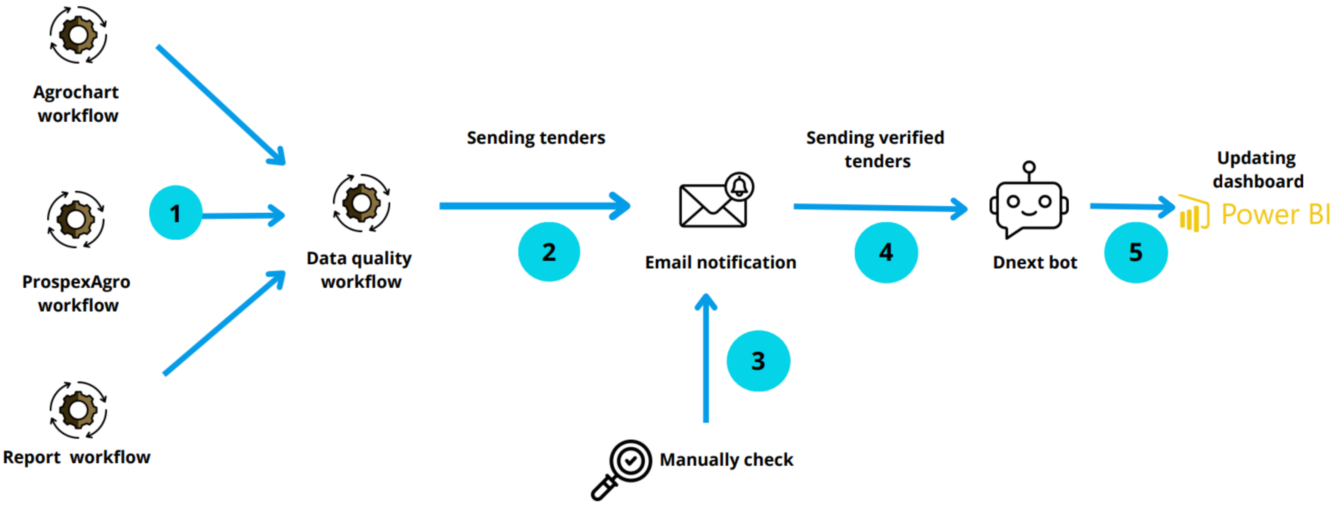


Figure 2.1: General Workflow of the Proposed Solution

- **Automation of Data Collection :**

  Our solution fully automates the manual process of data collection and extraction from multiple sources. Most data sources provide tender documents in text format, containing valuable information embedded within unstructured content. Our solution incorporates artificial intelligence techniques into the workflow of each data source to extract key information from tender text documents. This integrated approach eliminates the time-consuming task of manually gathering and processing data, while significantly reducing processing time and human error

- **Automation of Data Quality Process:**

  Data quality is automated through a series of predefined checks at various stages of the workflow. These automated checks are designed to identify and rectify inconsistencies, duplicates, and missing data. When discrepancies are detected in information from the same tender, the system automatically sends an email to the senior analyst, who is responsible for choosing between conflicting data points, and then processes the analyst's response automatically. The automated data quality workflow sends email notifications containing new information to the junior data analysts to make checking before updating the database, ensuring transparent communication and verification throughout the process

- **Client Value Enhancement Through Automation:**

  By deploying these automated workflows, we aim to eliminate delays, reduce errors and inaccuracies, and provide our clients with timely and accurate market insights through our dashboards.

2.3 Requirements Specification

This section outlines the analysis and specification of the project requirements, detailing the stakeholders involved and the functional and non-functional requirements of the solution.

2.3.1 Stakeholders

- The automated workflows are designed for data analysts responsible for collecting and analyzing tenders.
- The dashboard is designed for our clients to help improve their decision-making processes.

2.3.2 Functional Requirements

Our solution offers the following features:

- Capture only tender results or offers, excluding tender announcements.
- Capture and store all tenders published daily.
- Efficiently extract information from tender texts.
- Ensure the quality of data saved in the database.
- Automatically process decisions made by the senior data analyst in cases of inconsistent data and apply the necessary updates.
- Ensure dashboards are updated to facilitate decision-making based on the latest data.

2.3.3 Non-functional Requirements

Our solution must meet the following requirements:

- The solution should support the extensibility of adding new data sources.
- Daily tender updates must be processed and accessible on the dashboard within 12 hours of their publication.
- Availability: The platform should be available 99.99% of the time.

2.4 SCRUM Project Management

This project follows the SCRUM methodology to ensure efficient development and delivery of our tender monitoring solution. SCRUM provides an agile framework that enables iterative development, continuous feedback, and adaptive planning throughout the project lifecycle. The following sections outline our team structure, project backlog, and sprint organization.

2.4.1 Scrum Team Roles

We will outline our SCRUM team:

- **The Product Owner:** Mr. Pablo FIGUEROA defines the needs, priorities, and features, and leads the development team's activities.
- **The Scrum Master:** Mr. Farouk JAZIRI organizes meetings and ensures the proper application of the Agile methodology in this role.
- **The Development Team:** Ghaith KRIFA, Firas FEKIH and Ebtihel KANTAOUI perform the necessary work to deliver a product increment at the end of each iteration.

2.4.2 Product Backlog

The product backlog is a prioritized list of features and requirements that guide the development process. Each user story represents a functional need expressed from the perspective of an end user or stakeholder. The backlog is continuously refined to ensure the team focuses on delivering the highest value items first. Table 2.1 below presents the user stories:

ID	User Story	Priority	Complexity
1	As a data analyst, i receive tender news via email in a structured, efficient, and high-quality data format.	1	High
2	As a client, i have access to an updated dashboard that aids in enhancing my decision-making process.	2	High

Table 2.1: Product Backlog

2.4.3 Sprint Planning

The user stories defined in the product backlog are prioritized. After team discussions, we identified three sprints. Table 2.2 below provides details:

Sprint Number	**Description**
Sprint 1	Develop an automated workflow for each data source.
Sprint 2	Implement a data quality workflow.
Sprint 3	Design dashboards and deploy the solution on the server.

Table 2.2: Sprint Planning Overview

2.5 Work Environment

2.5.1 Hardware Environment

For the completion of my end-of-studies project, I used a laptop with the following specifications:

- Intel Core i5 Processor.
- 16 GB of RAM.
- 512 GB SSD.

2.5.2 Development Environment

- **LangChain:** is a framework for developing applications powered by LLMs. LangChain simplifies every stage of the LLM application lifecycle. [4]

Figure 2.2: LangChain Logo

- **Apache Airflow:** Apache Airflow is used for the scheduling and orchestration of data pipelines or workflows. Orchestration of data pipelines refers to the sequencing, coordination, scheduling, and managing of complex data pipelines from diverse sources. [5]

Figure 2.3: Apache Airflow Logo

- **Power BI:** Microsoft Power BI is a data visualization and reporting platform that is used by businesses and professionals. [6]

Figure 2.4: Microsoft Power BI Logo

- **Mongo DB:** MongoDB is a tool that can manage document-oriented information, store or retrieve information. MongoDB is used for high-volume data storage, helping organizations store large amounts of data while still performing rapidly. [7]

Figure 2.5: MongoDB Logo

- **Docker:** Docker is an open-source platform that enables developers to build, deploy, run, update and manage containers. [8]

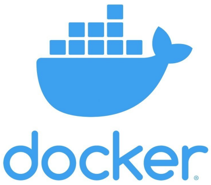


Figure 2.6: Docker Logo

2.6 Conclusion

In summary, this chapter has established the groundwork for the project by analyzing its context, challenges, and proposed solutions. It also defined detailed requirements, outlined the SCRUM-based management approach, and described the technical environment essential for effective project implementation.

Chapter

3

SPRINT 1: Data Sources Workflows Development

Contents

3.1 Introduction

In this chapter, we present a comprehensive overview of the workflows developed for each data source. We provide detailed explanations of these workflows, describe the AI models integrated within them, and examine the development and performance evaluation of these models.

3.2 Sprint Backlog

ID	User Story	Priority	Tasks
1.1	Develop a workflow to process and extract key information from recent tenders mentioned in the report.	1	- Design and plan the architecture of each workflow. - Develop AI models responsible for extracting key tender information. - Build a model to identify sentences indicating tender offers and results, and then standardize them into a uniform template. - Integrate and test all combined components of different workflows.
1.2	Develop a workflow to extract key information from recent tender texts provided by the AgroChart API.	2	
1.3	Develop a workflow to extract key information from recent tender texts provided by the ProspexAgro API.	3	

Table 3.1: Sprint Backlog

3.3 Workflow Report

This section provides an overview of the report workflow, highlighting the data source, key advantages, and challenges involved in processing South Korean tender reports. It also outlines the architecture and the sequential tasks that make up the workflow to ensure efficient data extraction, mapping, and loading.

3.3.1 Data Source

DNEXT has a partner based in South Korea who sends a weekly PDF report on South Korean tenders. Additionally, the report provides information about tenders for corn, feed wheat, and soybean meal.

Key Advantages:

- Easy to process as most key information is consistently located in the same sections of the PDF report.
- May include tenders not published by other sources.

Extracting Key Information Challenges:

- There are slight variations between PDF reports, requiring the extraction algorithm to accommodate multiple conditions to successfully capture all data.

3.3.2 Workflow Report Architecture

Figure 3.1 below provides our proposed solution for automating data collection from the PDF Report:

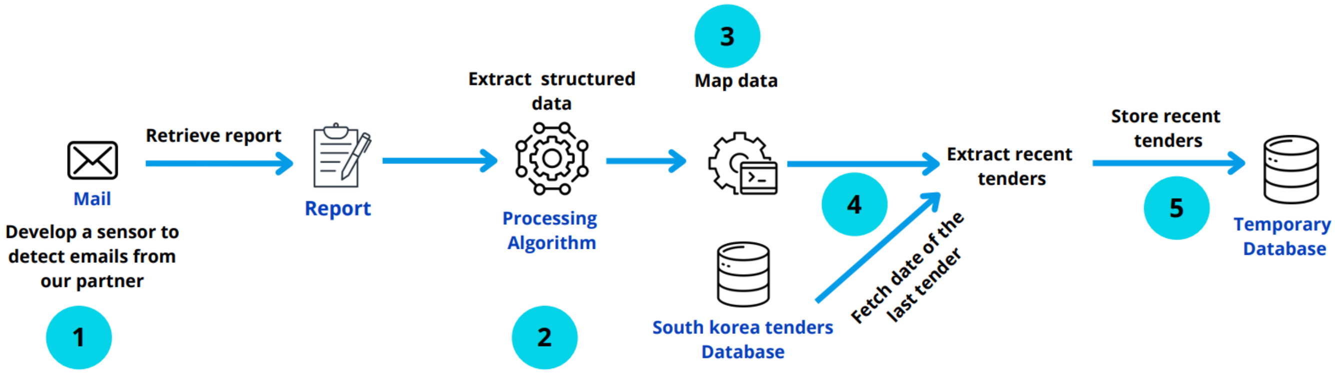


Figure 3.1: Workflow Report

To implement this architecture, we have developed a workflow consisting of four distinct tasks executed sequentially, each with a specific objective:

- **PDF Capturing:**

  In this task, we develop a custom sensor designed to detect the latest email from our partner. The sensor extracts the PDF attachment and loads it locally.

- **Data Processing:**

  This task involves using methods to extract the necessary information and format it according to our requirements.

- **Data Mapping:**

  Since data originates from multiple sources, standardization is essential to ensure consistency. For example, the US dollar might appear as "USD" and as "$" in another data source. Therefore, we must map and transform all data into standardized values to facilitate subsequent processing.

- **Data Loading:**

  During this task, we load the extracted tender information into a temporary database where it goes through data quality checks to ensure accuracy before being saved in the final database. Additionally, this task includes the deletion of the PDF from the local storage.

3.4 AgroChart Workflow

This section details the AgroChart workflow, focusing on the extraction and processing of tender data via the platform's API. It discusses the advantages and challenges of using this data source and presents the workflow architecture designed to address the challenges of extracting key information through AI models.

3.4.1 Data Source

AgroChart is a comprehensive online platform that provides real-time news, market data, and analysis related to global agricultural markets. We use this platform's API to obtain tender data.[9]

Key Advantages:

- AgroChart is the most consistent data source when it comes to agricultural market information.
- Most of the information we need is contained within the tender texts.
- AgroChart provides data on offers and results of tenders. We don't need extra processing to filter out tender announcements.

Extracting Key Information Challenges:

- This source provides tender data in text format, unlike the previous source, we need to process this text using AI models to extract the key information we require.

3.4.2 Agrochart Workflow Architecture

Figure 3.2 below provides our proposed solution for automating data collection from the Agrochart API:

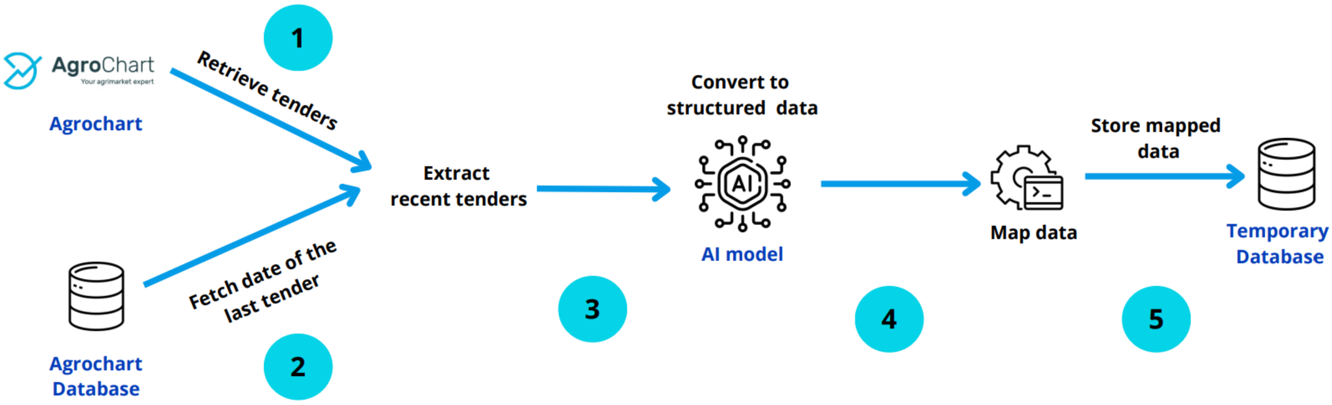


Figure 3.2: AgroChart Workflow

To implement this architecture, we have developed a workflow consisting of four distinct tasks executed sequentially, each with a specific objective:

- **New Tenders Extraction:**
  In this task, we aim to capture newly published tenders by checking the AgroChart database, where we store both the publication date and the tender text.
- **Data Processing:**
  In this phase, we implement the model responsible for extracting key information from the tender texts.
- **Data Mapping:**
  As mentioned previously, the mapping step transforms data into standardized formats to ensure uniformity.
- **Data Loading:**
  During this task, we load the extracted tenders information into a temporary database where it goes through data quality checks to ensure accuracy before being saved in the final database.

3.4.3 Model Development for Key Information Extraction

As mentioned in the data source section, AgroChart provides tenders in a structured text format. Before implementing the workflow, a model is needed to extract key information. This section outlines the architecture used for extraction and compares the performance of the models employed.

3.4.3.1 Information Extraction Approaches using LLMs

1. **Tool/Function Calling Mode**
   Some LLMs support a tool or function calling mode, enabling them to produce output strictly adhering to a predefined schema. This mode is generally the most reliable and easiest to integrate with software systems. It operates as follows:

- The user defines a precise schema or function signature detailing the expected structure and data types.
- The LLM generates output that exactly matches the defined schema, ensuring consistency and minimal errors.

2. **JSON Mode**

   In this mode, LLMs are prompted explicitly to produce valid JSON output. Unlike function calling, the schema is embedded within the prompt rather than enforced by the system. While the model strives to produce valid JSON, occasional syntax errors or deviations may occur. The process involves:

   - Providing instructions within the prompt to output data in JSON format.
   - The model attempts to comply, but generated output should be validated for correctness.

3. **Prompting-Based (Free-Text & Parsing)**

   This approach relies on instructing the LLM to generate text following a specified format, which is then parsed by downstream code using methods like regular expressions or custom logic. This method is applicable to any LLM but is the least reliable due to the variability in natural language generation. It works as follows:

   - The prompt describes the desired output format in free text.
   - The model produces a text response attempting to follow the instructions.
   - External parsing extracts the relevant data from the generated text.

Table 3.2 presents a benchmark comparing the different approaches for structured output generation:

Approach	**Benefits**	**Limitations**
Tool/Function Calling Mode	- Explicit schema validation by the model - High reliability of structured output	- Only available in models with function-calling support - May require additional API integration effort
JSON Mode	- Works with most LLMs even without tool support - More structured than free-text parsing	- May produce invalid or malformed JSON - Requires careful prompt engineering
Prompting-Based (Free-Text & Parsing)	- Compatible with virtually all LLMs - Flexible for custom formats	- Output may be inconsistent - Requires robust parsing and error handling

Table 3.2: Structured-output approaches comparison

3.4.3.2 Message Types for Prompt Engineering

LangChain provides several message types that play a crucial role in effective prompt engineering. These message types help structure conversations clearly and guide the AI's behavior effectively:

- **SystemMessage:**

  This type of message is sent by the system to give the AI model some basic instructions or context about how to respond during the conversation. It's like setting the ground rules for the AI. It's often the first message in a conversation.

- **HumanMessage:**

  This is the message sent by the user. It contains a question, command, or request. The

AI uses this message to understand what the user is asking or instructing it to do. The AI will process this input to generate an appropriate response.

- **AIMessage:**

  This is the response the AI generates based on the input from the user (HumanMessage) and any prior context. It could be the AI answering a question, providing clarification, or executing a task. If the AI calls an external tool or function, the response might also include results from that tool.

3.4.3.3 Model Architecture

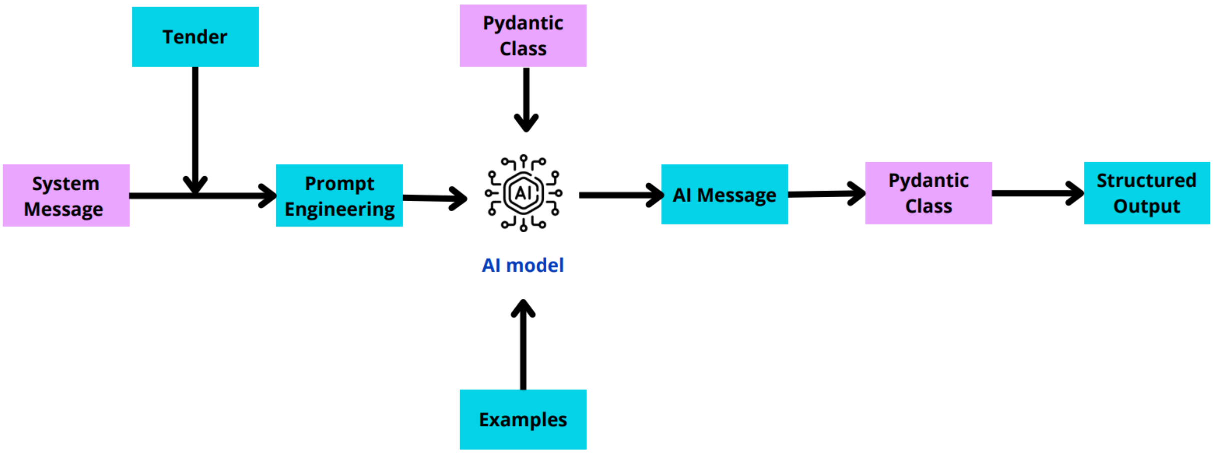


Figure 3.3: Model Architecture

We adopt the first approach: Tool/Function Calling mode. Figure 3.3 illustrates the architecture used to extract key information. Let's delve into this architecture to understand its components and how it operates.

- **Pydantic Class:**

  We begin by developing a Pydantic class, which defines the key information to be extracted. In our case, we aim to extract 15 key pieces of information. Each piece is described along with its data type (such as string, integer, or float). The Pydantic class provides the model with an overview of the details of the information we need to extract.

- **Prompt Template:**

  The prompt template guides LLM during the extraction process. Each of the 15 information fields has specific rules for extraction. In this process, we use the HumanMessage to pass the tender text to the LLM, ensuring it has the necessary data to work with. The SystemMessage is then used to provide the model with the specific rules for extracting each field. This combination of messages helps the LLM understand both the input data and the guidelines for accurate extraction.

- **Examples:**

  We also provide the LLM with real examples of text and the key information that the model is expected to extract. These examples help guide the LLM to understand the desired output. Various examples are included to give a general idea of the expected results.

3.4.3.4 Model's Performance

Our architecture is now ready for implementation. We tested it using models that support the tools function-calling mode. We selected three models: DeepSeek R1, Mistral Medium, and Mistral Small. Each model was tested with this architecture on 30 documents containing a total of 100 tenders. Table 3.3 below shows the accuracy of each model on these 100 tenders:

Field	Deepseek R1	Mistral Small	Mistral Medium
Number of tenders extracted in one text	95%	98%	100%
Reported date	100%	100%	100%
Country	100%	100%	100%
Product	100%	100%	100%
Quality	98%	93%	95%
Price	99%	95%	98%
Delivery start	95%	93%	96%
Delivery end	98%	98%	98%
Quote type	98%	93%	96%
Delivery term	95%	95%	95%
Company	95 %	90%	98%
Currency	100%	100%	100%
Unit	100%	100%	100%
Origin	100%	100%	100%
Tenderer	100%	100%	100%

Table 3.3: Extraction accuracy Comparison

When comparing the accuracies of the models across different attributes, we observed that their performances were generally similar, with only minor variations between them. However, DeepSeek R1 consistently showed lower performance in extracting all tenders from a single text compared to the other models. Ultimately, we selected Mistral Medium for this task because it achieved the best results in capturing all tenders within a single document. Since our primary goal is to extract all tenders comprehensively, any missing details can later be supplemented with information from other sources.

3.5 ProspexAgro Workflow

This section outlines the ProspexAgro workflow, focusing on extracting tender data from the Prospex Telegram channel. It highlights the advantages and challenges of using this data source and describes the workflow architecture designed to overcome the challenges.

3.5.1 Data Source

ProspexAgro is a commodity brokerage firm specializing in agricultural markets such as grains, pulses, and oils. We use the Telegram channel API of this platform to obtain tender data, enabling automated access to up-to-date information on agricultural tenders. [10]

Key Advantages:

- This source is the first to publish tender data, sometimes up to 4-5 days ahead of other sources, which is advantageous as we need to capture tenders as soon as they are available.

Extracting key Information Challenges:

- The Prospex channel provides various information related to agriculture but does not focus exclusively on tenders, so we must filter the data to include only tender-related content.
- The channel publishes diverse tender-related content such as announcements, offers, and results, which adds complexity to the extraction process. However, our business specifically focuses on extracting only offers and results.
- Tender data is presented in an unstructured text format, meaning there is no fixed template and the format varies with each publication.

3.5.2 ProspexAgro Workflow Architecture

Figure 3.4 provides our proposed solution for automating data collection from the ProspexAgro Telegram channel API:

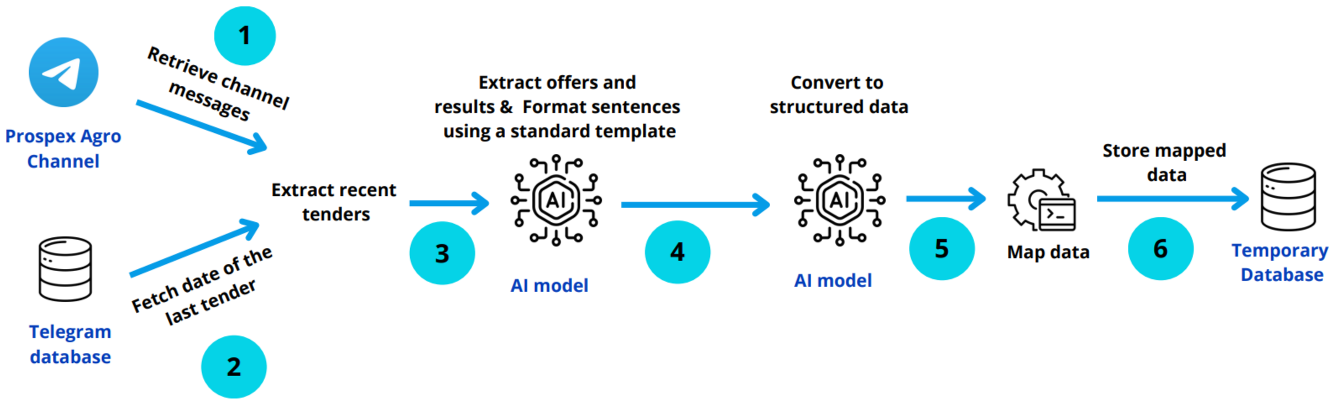

1
Retrieve channel messages
Prospex Agro Channel
Telegram database
Fetch date of the last tender
2
Extract recent tenders
3
Extract offers and results & Format sentences using a standard template
AI model
4
Convert to structured data
AI model
5
Map data
Store mapped data
6
Temporary Database

Figure 3.4: ProspexAgro Workflow

To implement this architecture, we have developed a workflow consisting of four distinct tasks executed sequentially, each with a specific objective:

- **Channel Message Capturing and Filtering for Newly Published Tenders:**

  This task focuses on capturing newly published tenders by comparing message dates. We retrieve messages published after the last recorded date and filter to keep only those containing tenders. Since each message clearly indicates its topic at the top, AI models are not required for filtering.

- **Model Implementation for Excluding Announcements and Standardizing Tender Data:**

  In this phase, we implement a model that filters tender offers and results while excluding announcements. Additionally, the model structures and standardizes the published tender text format to prepare it for accurate extraction of key information.

- **Data Processing:**

  In this phase, we implement the model responsible for extracting key information, followed by mapping this information to a standardized format.

- **Data Loading:**

  During this task, we load the extracted tenders information into a temporary database where it goes through data quality checks to ensure accuracy before being saved in the final database.

3.5.3 Model for Filtering and Structuring Tender Text

As mentioned in the description of this data source, we face two main challenges before extracting keywords. To address these challenges, we developed a prompt engineering approach. The prompt clearly defines what constitutes tender offers and results, enabling the extraction of only the relevant sentences while excluding announcements. This is not a text classification task, as a single message may contain a mix of offers, results, and announcements. The prompt also includes a standardized template to restructure the extracted content into a consistent textual format.

Figure 3.5 below illustrates how we convert channel messages into structured tender text.The system message guides the LLM by providing detailed instructions and examples on how to identify and extract only the offers and results, and how to reformat them using the predefined template.

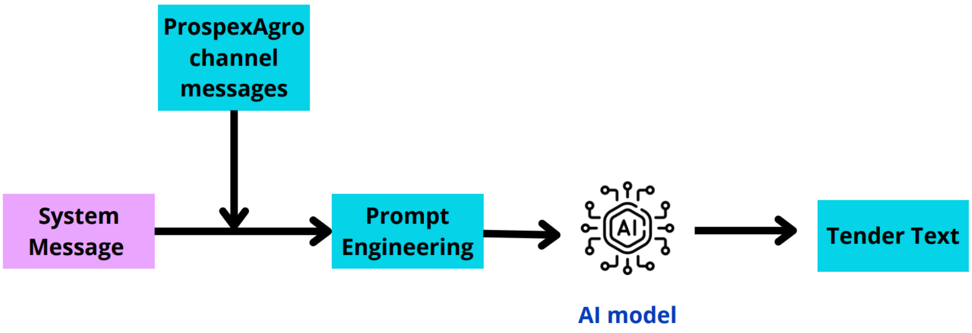


Figure 3.5: Tender Text Extraction Approach

3.5.3.1 Model's Performance

The prompt engineering is now ready for implementation.We selected two models: DeepSeek R1, Mistral Medium and tested each using prompt engineering on a set of 40 messages:

- 15 messages containing announcements, offers, and results simultaneously.
- 10 messages containing only announcements.
- 15 messages containing results or a combination of offers and results. These 15 messages include a total of 35 tenders.

Table 3.4 below shows a comparison of the performance metrics between Deepseek R1 and Mistral Medium:

Metric	Deepseek R1	Mistral Medium
Including sentences that indicate offers/results of tenders	99%	99%
Total information supposed to be mentioned in output text	95%	93%
Filtering out tenders announcement	100%	100%
Average speed of answer	20 seconds	5 seconds

Table 3.4: Performance Metrics Comparison

The two models demonstrated high performance across all the metrics we evaluated, with only slight differences in the total amount of information extracted from the text. However, we decided to implement Mistral Medium due to its faster response time compared to Deepseek R1. Speed is a critical factor for us, as it helps minimize task failures caused by sensitivity to internet connection speed.

3.6 Conclusion

In conclusion, this chapter has provided a detailed exploration of automated workflows specifically designed for managing data sources. The chapter outlined the architecture of these workflows and highlighted the tasks involved in processing and integrating data from various origins. Additionally, we discussed the integration of LLMs into these workflows, detailing their development process and analyzing their performance to enhance data processing capabilities. By understanding these automated systems and the role of LLMs within them, we are now well-prepared to transition to the next critical phase. This next phase involves ensuring the quality of data before it is saved in the final database.

Chapter

4

SPRINT 2: Data Quality Workflow Developement

Contents

4.1 Introduction

Our data source workflows are now ready to feed data into our database. However, since tender data arrives with varying delays across different sources, it is essential to verify its quality before storing it in the final database.

In this chapter, we provide a detailed explanation of the data quality aspects, describe the workflow development, outline the handling of each step, and detail the involvement of senior experts in ensuring data quality.

4.2 Sprint Backlog

ID	User Story	Priority	Tasks
2.1	Design the data quality workflow.	1	- Develop and validate the tender matching and rating algorithms to identify the highest quality information. - Integrate and test all combined components of this workflow. - Define and implement conditions to control the execution of data sources workflows.
2.2	Develop the tender matching algorithm.	2	
2.3	Develop the rating algorithm for selecting the highest quality information.	3	
2.4	Implement conditional logic for executing the data sources workflows.	3	

Table 4.1: Sprint Backlog

4.3 Data Quality Key Checks

Ensuring high data quality is crucial for accurate analysis and decision-making in tender processing. Our workflow involves multiple verification steps to validate and consolidate information from various sources. These checks help to identify inconsistencies, fill missing data, and prioritize the most reliable information.

- **Checking for complete information:**

  Our initial goal is to ensure that all information is complete without missing fields. By combining multiple data sources, we aim to fill in all 14 information fields completely.

- **Verifying price, delivery end, and quantity:**

  Sometimes conflicting information appears in these fields across different sources. In such cases, a senior analyst with domain expertise steps in to make the correct determination.

- **Prioritizing information with CNF delivery term:**

  We prioritize information with CNF delivery term. While some tenders may be published with other terms like FOB, we give preference to CNF data whenever it is available.

- **Prioritizing prices in US dollars and quantities in metric tons:**

  Prices may sometimes be reported in various currencies, such as euros or cents. When price information is quoted in US dollars, we prioritize it and automatically update the corresponding quantity to metric tons.

4.4 Data Quality Workflow Architecture

Figure 4.1 provides our proposed solution for automating key checks Telegram channel API:

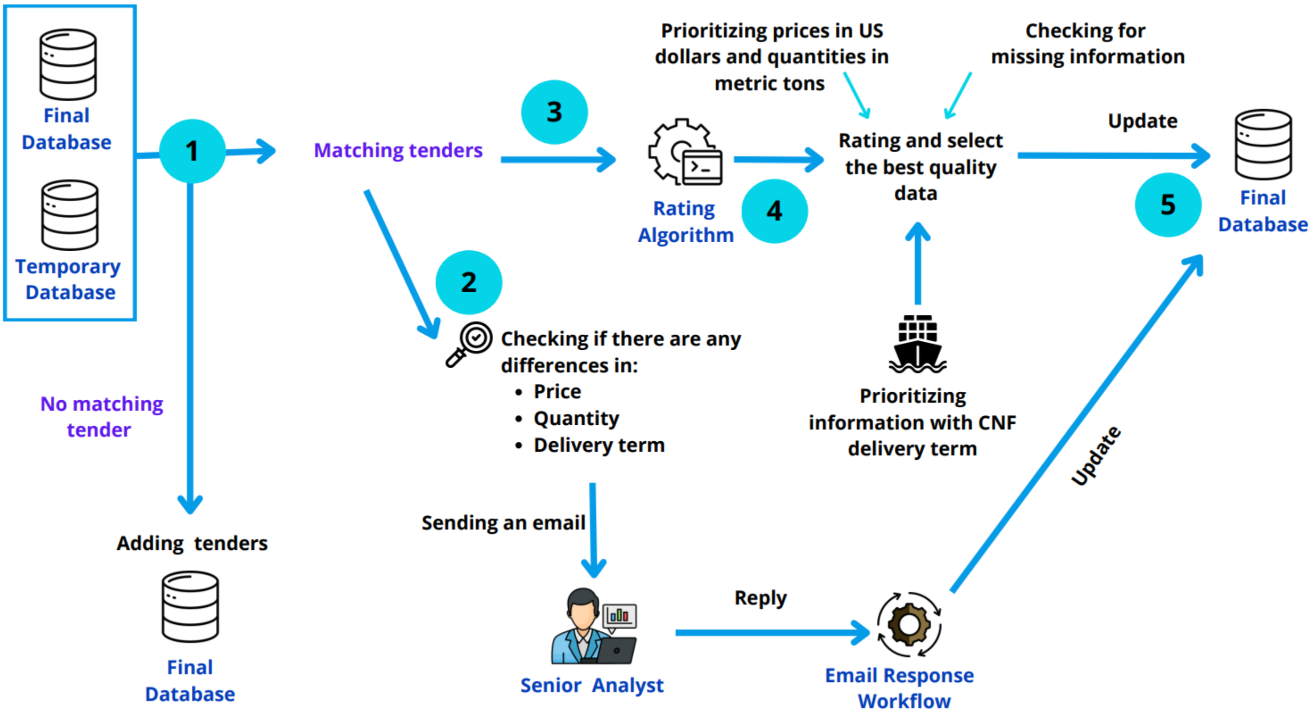


Figure 4.1: Data Quality Workflow

To implement this architecture, we have developed a workflow consisting of five distinct tasks executed sequentially, each with a specific objective:

- **Matching Tender:**

  Extract tenders from the temporary database and try to match them with tenders saved in the last six days. We assign each field a weight according to its importance:

  For two tenders $T^{(1)}$ and $T^{(2)}$, we define the field-wise similarity $\mathrm{sim}_f(T^{(1)}, T^{(2)})$. Then the overall similarity is

$$\mathrm{Sim}(T^{(1)}, T^{(2)}) = \frac{\sum_f w_f \,\mathrm{sim}_f(T^{(1)}, T^{(2)})}{\sum_f w_f}. \tag{4.1}$$

We choose a threshold $\theta = 0.85$. Two tenders are considered a match if

$$\mathrm{Sim}(T^{(1)}, T^{(2)}) \;\geq\; \theta\,. \tag{4.2}$$

Table 4.2 below presents the weight assignments for each tender field:

Field	**Weight (w_f)**
Date	0
Country	4
Tenderer	4
Product	4
Quality	1
Company	4
Origin	1
Delivery Start	1
Delivery End	1
Quantity	1
Price	1
Currency	0
Unit	0
Delivery Term	1.5
Quote Type	2

Table 4.2: Tender Fields Weights

- **Adding New Tenders:**

  If no matching tenders are found, the tenders in the temporary database are considered new and are automatically saved to the final database.

- **Verifying Specific Field Values:**

  After matching, key fields like price, quantity, and delivery date are checked for consistency. If discrepancies are found, the senior analyst makes the final decision based on their experience.

- **Rating Matched Tenders and selecting the best quality data:**

  For each pair of matched tenders, a rating algorithm is applied to identify the highest quality data. If one tender's price is in USD, it will replace the price in the best-quality data, regardless of the original data quality. The same logic applies to delivery term.

Missing information in the best-quality data is also supplemented with data from the matched tender.

- **Updating Tenders:**

  Updates are applied after determining the highest quality data.

4.5 Rating Algorithm

- **Required Fields Score:**

  The algorithm first verifies the presence of mandatory fields: date, country, product, price, delivery end, quantity, and delivery term. If any required field is missing, the record receives a required score of 0. Otherwise, if all required fields are present and valid, the score is 7.

- **Non-Required Fields Score:**

  Next, the algorithm evaluates optional fields. Each non-required field with valid data contributes 1 point to the non-required score.

- **Duplicate Value Penalty:**

  The algorithm checks for duplicate values among all non-null fields. If duplicates exist, it subtracts 1 point as a penalty to redundant data.

- **Final Score Calculation:**

  The total score is calculated as:

  $$FinalScore = RequiredScore + NonRequiredScore - DuplicatePenalty$$

  This balances the importance of required data completeness, rewards additional information, and penalizes redundancy to reflect overall data quality.

4.6 Auto-Detecting Decisions from Replies

As we previously discussed, there are three specific fields within the tender data that have the potential to contain Conflicts. To ensure accuracy and maintain data integrity, these Conflicts must be carefully reviewed and resolved by the senior analyst. Our senior analyst will be responsible for evaluating these inconsistencies and deciding on the most appropriate course of action for each case.

To facilitate this review process, an automated email notification system has been put in place. Whenever a discrepancy is detected between the tender fields, an email will be sent directly to the senior analyst. This email is designed to provide a clear and comprehensive overview of the relevant information, enabling the analyst to make a well-informed decision.

The email includes:

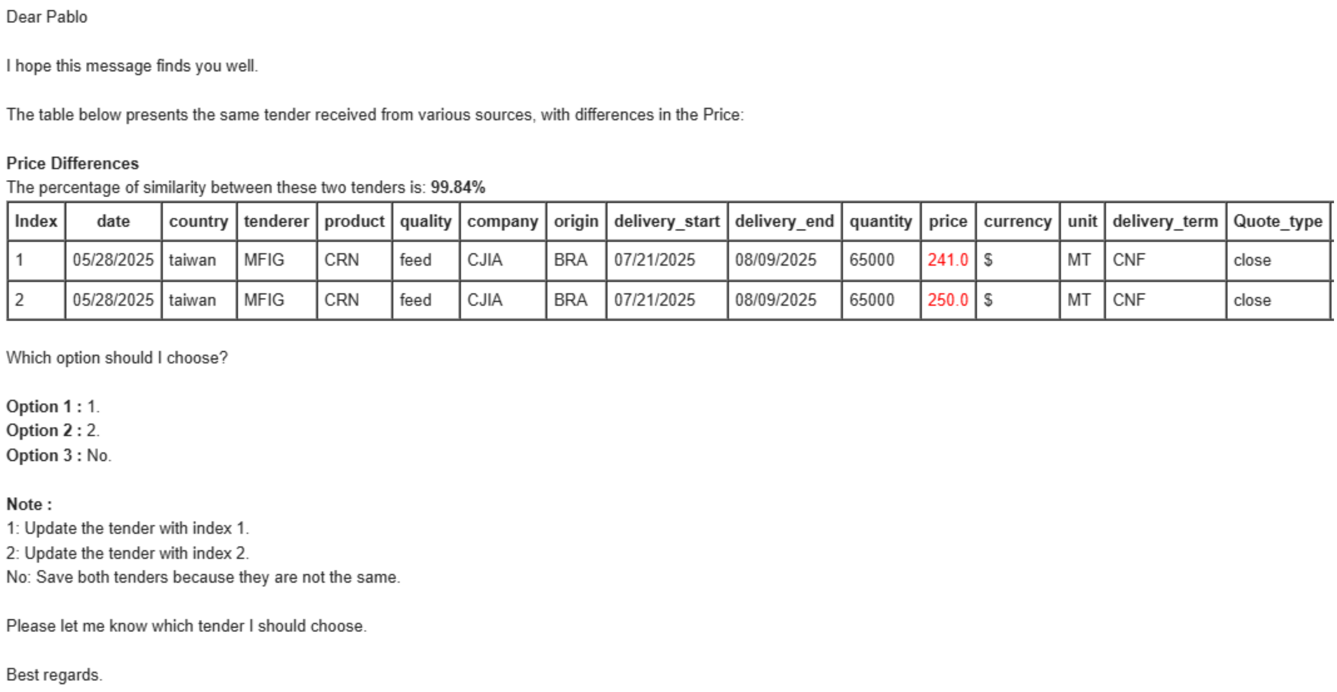
- A table displaying the tender details.
- Differences are highlighted in red to stand out.
- The similarity percentage, which indicates how closely the tenders match. This percentage is important because, despite some tenders appearing similar, they may still be considered different.

The email will also provide the following options for the senior analyst to choose from:

- Update the tender with index 1.
- Update the tender with index 2.
- Save both tenders as they are not identical.

Each discrepancy notification will follow this structure, enabling the senior analyst to make an informed decision based on the data provided.

Figure 4.2 illustrates an example of an email sent to a senior analyst, highlighting discrepancies in the price information.



Dear Pablo

I hope this message finds you well.

The table below presents the same tender received from various sources, with differences in the Price:

Price Differences
The percentage of similarity between these two tenders is: **99.84%**

Index	date	country	tenderer	product	quality	company	origin	delivery_start	delivery_end	quantity	price	currency	unit	delivery_term	Quote_type
1	05/28/2025	taiwan	MFIG	CRN	feed	CJIA	BRA	07/21/2025	08/09/2025	65000	241.0	$	MT	CNF	close
2	05/28/2025	taiwan	MFIG	CRN	feed	CJIA	BRA	07/21/2025	08/09/2025	65000	250.0	$	MT	CNF	close

Which option should I choose?

Option 1 : 1.
Option 2 : 2.
Option 3 : No.

Note :
1: Update the tender with index 1.
2: Update the tender with index 2.
No: Save both tenders because they are not the same.

Please let me know which tender I should choose.

Best regards.

Figure 4.2: Email Template

4.7 Analyst Decision Processing

The senior analyst will respond individually to each request. We have developed a workflow to automate the decision-making process and implementing the necessary updates. Figure 4.3 below illustrates the architecture of this workflow.

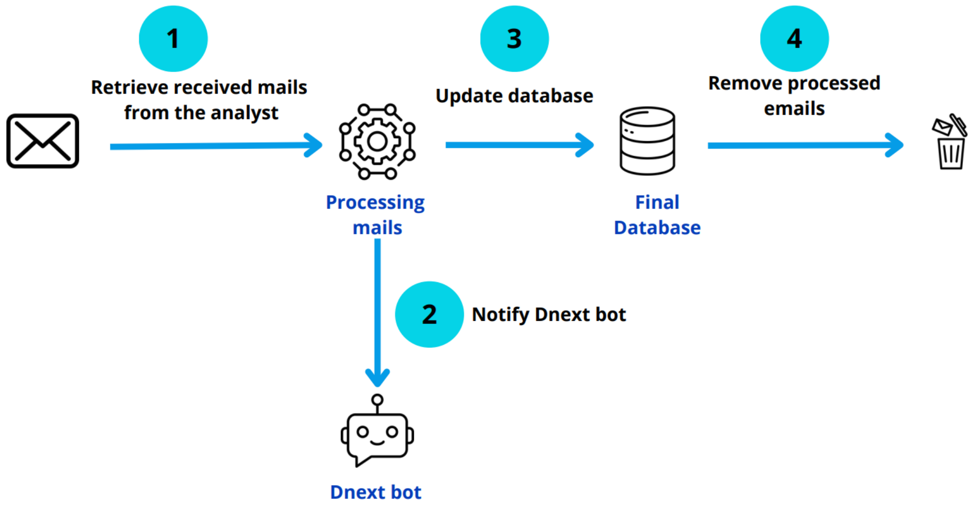


Figure 4.3: Analyst Decision Workflow Architecture

We have implemented a four-step sequential workflow to realize the architecture described:

- **Capture Email Replies:**

  Extract the senior analyst's responses from the inbox file.

- **Process Data:**

  Analyze the decisions provided by the senior analyst.

- **Load Result:**

  Update the final database according to the analyst's decisions.

- **Notify Dnext Bot:**

  Send the update summary to the Dnext bot, which then applies the necessary changes to its database.

4.8 Improvements of Data Sources Workflows

The development of the Data Quality workflow has enabled us to introduce significant improvements to the data source workflows. We have implemented an execution condition for these workflows that verifies whether the senior analyst has responded to all pending emails. This check is crucial because if the analyst has not made decisions on older tenders, uploading new tenders without resolving prior issues could cause serious problems. Therefore, before any data source workflow runs, we ensure that all previous tenders have been saved to the final database with the highest data quality. Figure 4.4 below illustrates the condition execution:

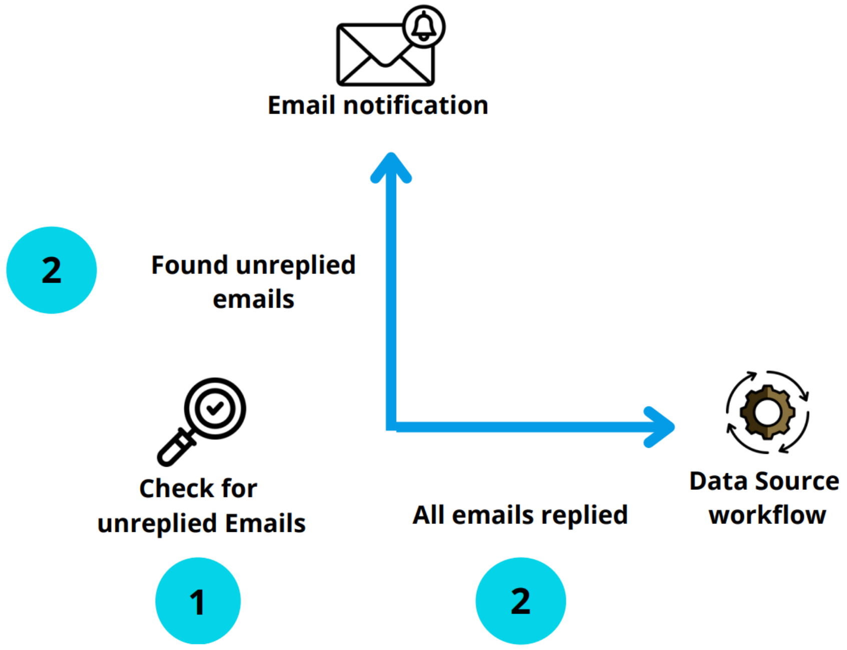
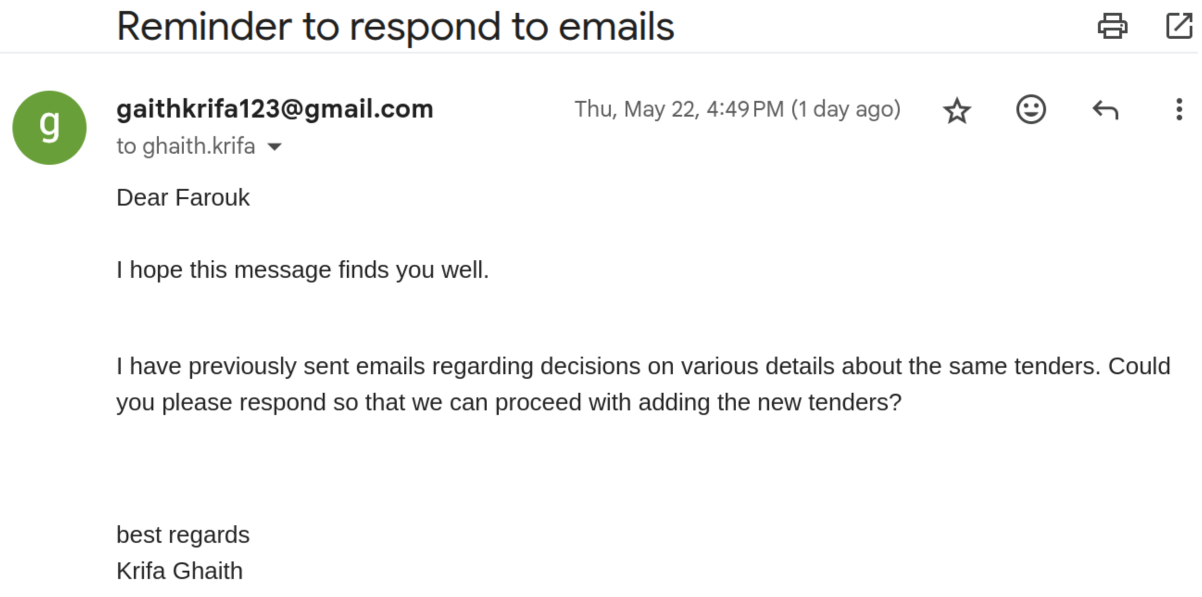


Figure 4.4: Condition Execution

Figure 4.5 presents an example of a reminder sent to the analyst to prompt replies on outstanding emails necessary for extracting new tenders:



Reminder to respond to emails

gaithkrifa123@gmail.com
to ghaith.krifa

Thu, May 22, 4:49 PM (1 day ago)

Dear Farouk

I hope this message finds you well.

I have previously sent emails regarding decisions on various details about the same tenders. Could you please respond so that we can proceed with adding the new tenders?

best regards
Krifa Ghaith

Figure 4.5: Email Reminder

4.9 Conclusion

In conclusion, this chapter introduced a robust automated data quality workflow aimed at ensuring the reliability and accuracy of tender data before its final storage. Through the implementation of automated tasks, all critical checks were systematically performed, guaranteeing the integrity and high quality of the data inserted into the final database.

Chapter

5

SPRINT 3: Deployment and Visualisation

Contents

5.1 Introduction

This chapter covers the project's final stage: deploying the developed workflows and creating dashboards. First, we explain how to deploy Apache Airflow DAGs using Docker, a containerization platform that ensures consistency, portability, and scalability, making it ideal for production environments. Next, we describe how to build interactive Power BI dashboards, enabling stakeholders to derive meaningful insights from the processed tenders.

5.2 Sprint Backlog

ID	User Story	Priority	Tasks
1.1	Create a Docker image for workflows to enable deployment on the server.	1	- Schedule the execution dates for the workflows. - Implement error handling for workflows with email notifications on failure. - Test Docker image deployment and workflow execution on the server.
1.2	Develop the dashboards.	2	- Design and implement dashboards. - Add interactive features for user engagement.

Table 5.1: Sprint Backlog

5.3 Solution Deployment

5.3.1 Docker Concepts

Before diving into the deployment steps, it is essential to understand some fundamental Docker concepts, particularly images, containers, and Dockerfile as they are at the core of how Docker operates and are the main elements used in deploying our workflows.

- **Containers:**

  A container is an isolated environment in which an application runs with its own configuration and dependencies. Containers can host various types of applications, such as websites, APIs, and databases.

  Each container is an instance of a Docker image and has its own runtime environment, including its own file system.

- **Images:**

  A Docker image defines the execution environment of containers. Just as classes are blueprints for objects in object-oriented programming, images are blueprints for containers.

  An image includes everything needed to run an application: directory structure, libraries, internal and external port bindings, and startup commands. Multiple containers can be created from a single image.

- **Dockerfile:**

  A Dockerfile is a script that specifies the instructions needed to build a Docker image. It is read from top to bottom during the build process and defines each step of the image creation, including the base image, dependencies, files to copy, and commands to execute.

Figure 5.1 below illustrates the core steps involved in using Docker: it starts with a Dockerfile, which contains instructions for building an image. This image is then used to launch one or more containers.[11]

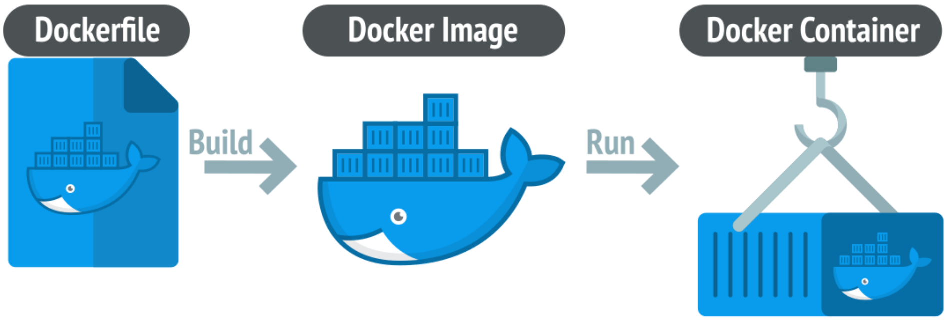


Figure 5.1: Containerization Steps

5.3.2 Workflows Deployment Steps

Below, we describe the key steps to deploy our workflows via Docker and configure Apache Airflow for automated email notifications, enabling analysts to respond efficiently to discrepancy requests.

a) Analyst Email Setup

We created a dedicated email account for analysts to handle communications regarding discrepancy requests. This address is integrated into the workflow so that Airflow can send automated notifications whenever analyst attention or action is required.

b) Docker Environment Preparation

To ensure reproducibility and consistency, we built a custom Docker image based on the official Apache Airflow image. Starting from this base, we installed additional system packages (e.g., Git) to support workflow dependencies and version control. We also listed all Python packages required by our workflows in a requirements file, which was used to install those dependencies inside the container.

c) Docker Image Building

After preparing the Dockerfile with all required instructions, we built the Docker image. This process packages the operating system, installed packages, Airflow components, and all workflow dependencies into a single image ensuring a consistent environment across deployments.

d) Airflow Container

Once the image was built successfully, we launched a Docker container from it. This container runs both the Airflow scheduler and web server, orchestrating workflow execution. Running Airflow in a container ensures isolation and simplifies dependency and runtime management.

e) Airflow Email Settings Configuration

To enable automated email notifications, we configured Airflow's SMTP settings including, the SMTP server address, port, and authentication credentials for the dedicated analyst email account. With these settings in place, Airflow can send alert emails directly from the workflows.

f) Workflows Testing

Before deploying to production, we conducted thorough testing by running the workflows inside the Docker container. This verified that all tasks executed correctly, dependencies were resolved, and email notifications were sent as expected. Testing helped identify and fix any issues with task scheduling, error handling, or email communication, ensuring reliability and smooth operation.

5.4 Workflow Monitoring and Management

After creating DAGs, Apache Airflow provides advanced features for workflow monitoring and management, such as real-time task execution status, visualization of task dependencies, error handling with retries, and the ability to dynamically adjust schedules based on operational needs. In our setup, we have configured the following settings:

- **Error Handling and Retries:**

  We configured a 3-minute retry delay before automatically restarting failed tasks. This approach handles minor errors effectively and ensures smooth execution of the overall workflow. If a task fails again after the retry, Airflow sends an email detailing the encountered error.

- **Scheduled Execution Calendar:**

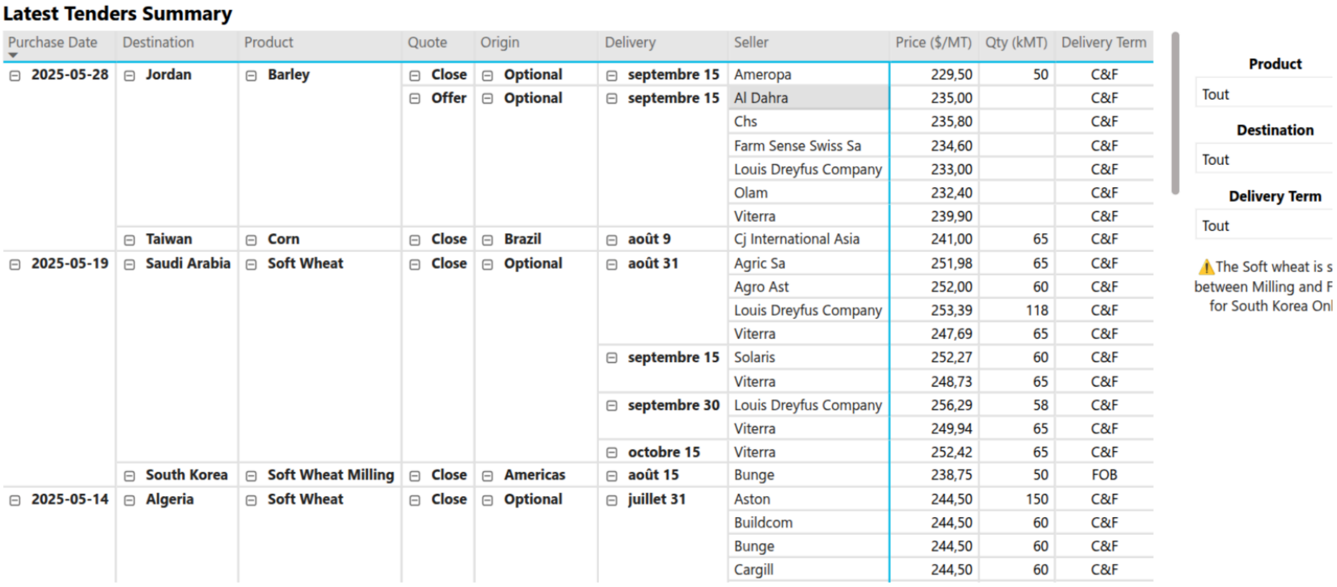
  - AgroChart workflow runs twice daily at 11:00 AM and 4:00 PM.
  - Prospex workflow runs twice daily at 11:30 AM and 3:30 PM.
  - PDF Report workflow runs once a week, every Monday at 11:00 AM.
  - Data Quality workflow is triggered whenever one of the data source workflows is executed.
  - Analyst Decision workflow runs every two hours throughout the day.

5.5 Power BI Report

In this section, we present a comprehensive review of the final Power BI report. We explore each page in detail, highlighting the insightful charts, interactive elements, and dynamic features. These pages include sections such as Tender Table, Tender Matrix, Quantities by Tender, and more. Each page is examined meticulously to reveal the analytical features and interactive functionalities that enhance the user experience. Across all pages, the two main axes of analysis are purchase date and delivery date.

5.5.1 Tender Table

This dashboard presents a comprehensive overview of offered and closed tenders, organized by purchase date or by delivery date. It displays essential information for each tender such as destination, product type, and origin. The table is designed to facilitate quick comparison and monitoring of tender activity over time. To enhance analytical flexibility, the page includes three interactive filters: Destination, Delivery Term, and Product Type. These filters allow users to customize the view based on specific criteria, making it easier to focus on relevant tenders. Figure 5.2 presents an example of a tender table dashboard:



Latest Tenders Summary

Purchase Date	Destination	Product	Quote	Origin	Delivery	Seller	Price ($/MT)	Qty (kMT)	Delivery Term
2025-05-28	Jordan	Barley	Close	Optional	septembre 15	Ameropa	229,50	50	C&F
			Offer	Optional	septembre 15	Al Dahra	235,00		C&F
						Chs	235,80		C&F
						Farm Sense Swiss Sa	234,60		C&F
						Louis Dreyfus Company	233,00		C&F
						Olam	232,40		C&F
						Viterra	239,90		C&F
	Taiwan	Corn	Close	Brazil	août 9	Cj International Asia	241,00	65	C&F
2025-05-19	Saudi Arabia	Soft Wheat	Close	Optional	août 31	Agric Sa	251,98	65	C&F
						Agro Ast	252,00	60	C&F
						Louis Dreyfus Company	253,39	118	C&F
						Viterra	247,69	65	C&F
					septembre 15	Solaris	252,27	60	C&F
						Viterra	248,73	65	C&F
					septembre 30	Louis Dreyfus Company	256,29	58	C&F
						Viterra	249,94	65	C&F
					octobre 15	Viterra	252,42	65	C&F
	South Korea	Soft Wheat Milling	Close	Americas	août 15	Bunge	238,75	50	FOB
2025-05-14	Algeria	Soft Wheat	Close	Optional	juillet 31	Aston	244,50	150	C&F
						Buildcom	244,50	60	C&F
						Bunge	244,50	60	C&F
						Cargill	244,50	60	C&F

Product: Tout

Destination: Tout

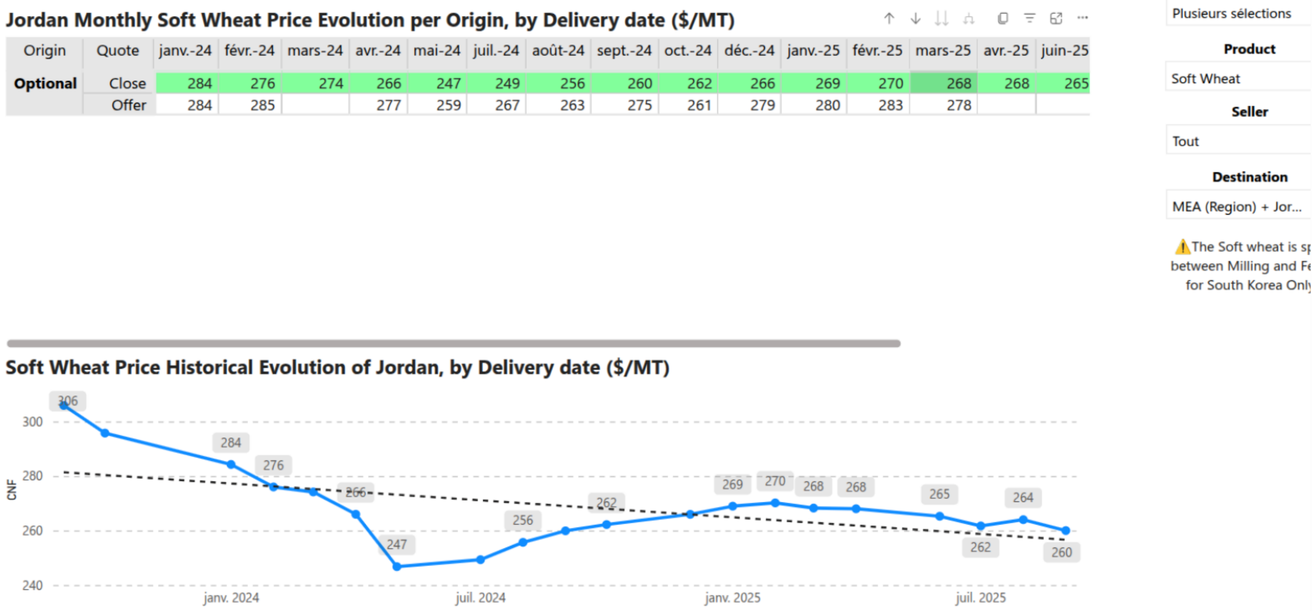
Delivery Term: Tout

⚠The Soft wheat is s between Milling and F for South Korea Onl

Figure 5.2: Tender Table Dashboard

5.5.2 Tender Matrix

This dashboard provides an interactive overview of prices for selected countries and products. Users can filter by product type, seller, and destination country or region. The price matrix table displays the monthly price evolution for the chosen product by origin, showing how prices vary across different months and sources. This helps users compare price differences for specific products in their selected country. The historical price trend graph presents the long-term price evolution of the selected product in the chosen country or region, highlighting trends, fluctuations, and seasonality. These visuals support customized market analysis and informed procurement and pricing decisions based on the selected filters. Figure 5.3 presents an example of a tender matrix dashboard:



Jordan Monthly Soft Wheat Price Evolution per Origin, by Delivery date ($/MT)

Origin	Quote	janv.-24	févr.-24	mars-24	avr.-24	mai-24	juil.-24	août-24	sept.-24	oct.-24	déc.-24	janv.-25	févr.-25	mars-25	avr.-25	juin-25
Optional	Close	284	276	274	266	247	249	256	260	262	266	269	270	268	268	265
	Offer	284	285		277	259	267	263	275	261	279	280	283	278		

Plusieurs sélections

Product

Soft Wheat

Seller

Tout

Destination

MEA (Region) + Jor...

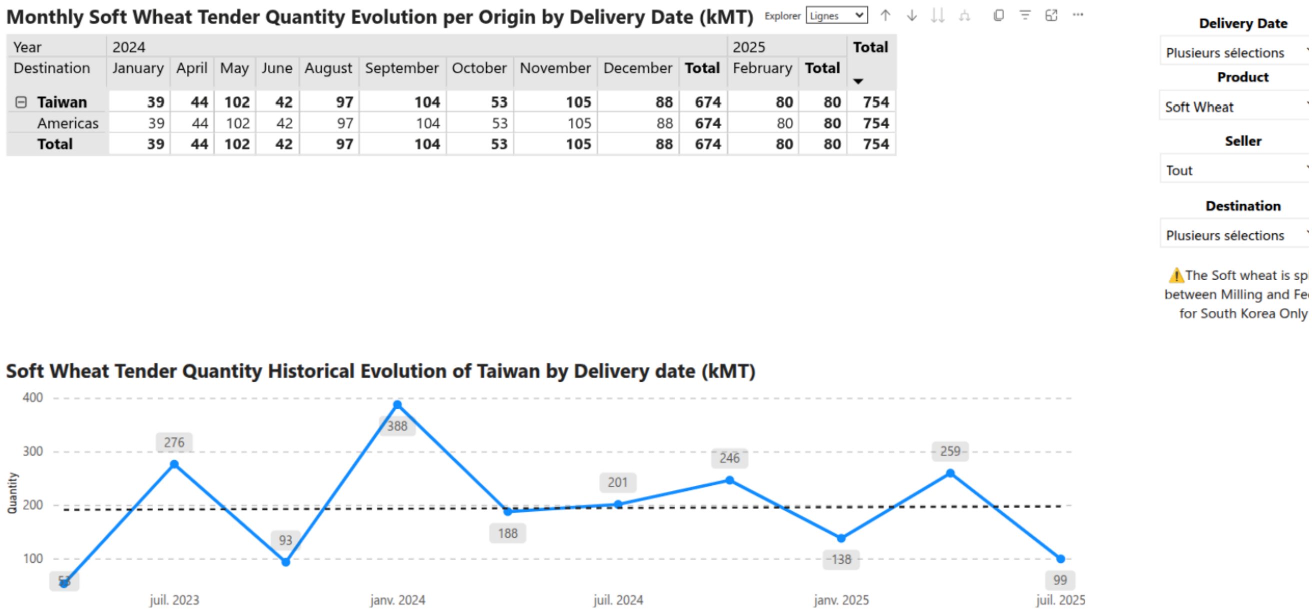
⚠The Soft wheat is sp
between Milling and Fe
for South Korea Only

Soft Wheat Price Historical Evolution of Jordan, by Delivery date ($/MT)

Figure 5.3: Tender Matrix Dashboard

5.5.3 Quantities by Tender

This dashboard provides a comprehensive overview of commodity tender quantities, allowing users to analyze procurement patterns across different origins and time periods. Users can filter data by product specifications, seller, and destination to customize their analysis. The monthly tender matrix displays procurement volumes by origin country and delivery month, enabling users to identify seasonal patterns, peak supply periods, and major supplier contributions. This view also reveals procurement activities and volume variations across suppliers and time frames. The historical evolution chart tracks long-term trends in tender quantities, illustrating market dynamics. Figure 5.4 presents an example of quantitites by tender dashboard:



Monthly Soft Wheat Tender Quantity Evolution per Origin by Delivery Date (kMT)

Year	2024										2025		Total
Destination	January	April	May	June	August	September	October	November	December	Total	February	Total	
⊟ **Taiwan**	**39**	**44**	**102**	**42**	**97**	**104**	**53**	**105**	**88**	**674**	**80**	**80**	**754**
Americas	39	44	102	42	97	104	53	105	88	**674**	80	**80**	**754**
Total	**39**	**44**	**102**	**42**	**97**	**104**	**53**	**105**	**88**	**674**	**80**	**80**	**754**

Delivery Date: Plusieurs sélections

Product: Soft Wheat

Seller: Tout

Destination: Plusieurs sélections

⚠️The Soft wheat is sp
between Milling and Fe
for South Korea Only

Soft Wheat Tender Quantity Historical Evolution of Taiwan by Delivery date (kMT)

Figure 5.4: Quantities by Tender Dashboard

5.5.4 Purchase History

This dashboard provides a detailed analysis of the commodity purchase history for a specific country and crop year. It displays the relationship between procurement quantities, individual purchase prices, and weighted average pricing over time. The visualization combines three key metrics on a dual-axis chart: procurement quantities shown as blue bars, individual purchase prices represented by the purple line, and the weighted average price displayed as a dotted black line. This multi-dimensional view enables users to understand how purchasing volumes correlate with price movements and market conditions. The timeline spans the entire crop year period, allowing users to track seasonal procurement patterns and price evolution. Figure 5.5 presents an example of purchase history dashboard:

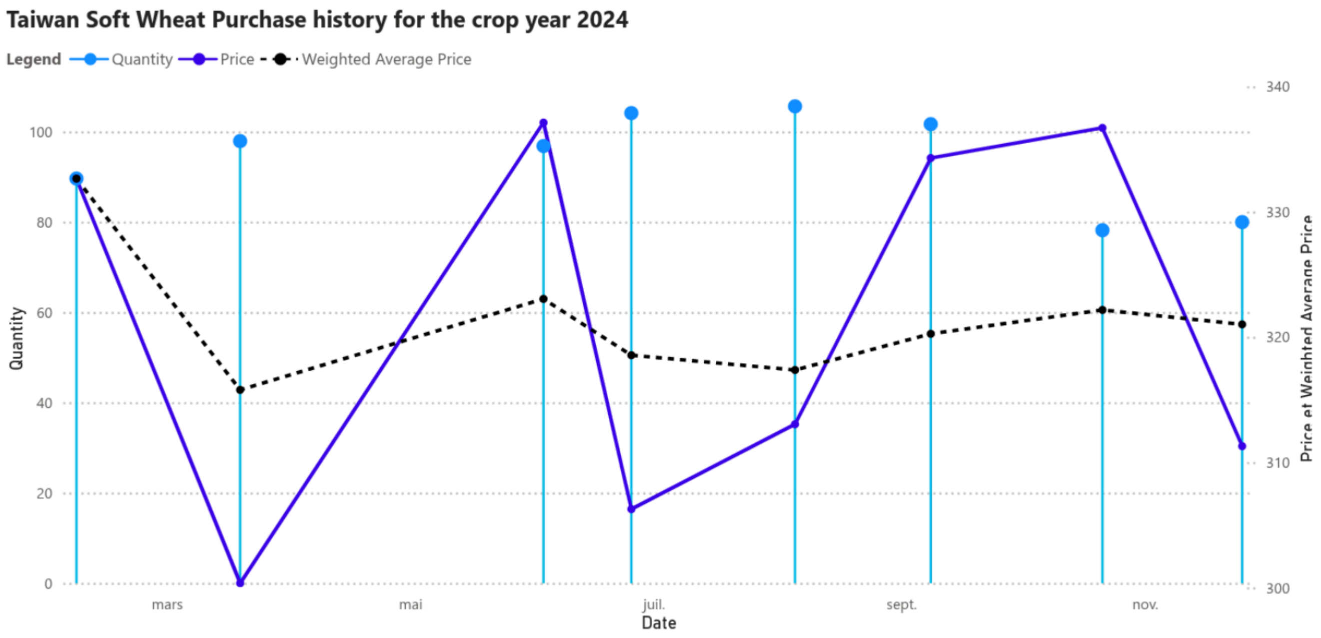


Figure 5.5: Purchase History Dashboard

5.6 Case Study: Analysis of Soft Wheat Purchasing Strategies

This section presents a case study on the purchasing strategies of Egypt and Tunisia for soft wheat.

5.6.1 Analysis of Egypt's 2023 Purchasing Strategy

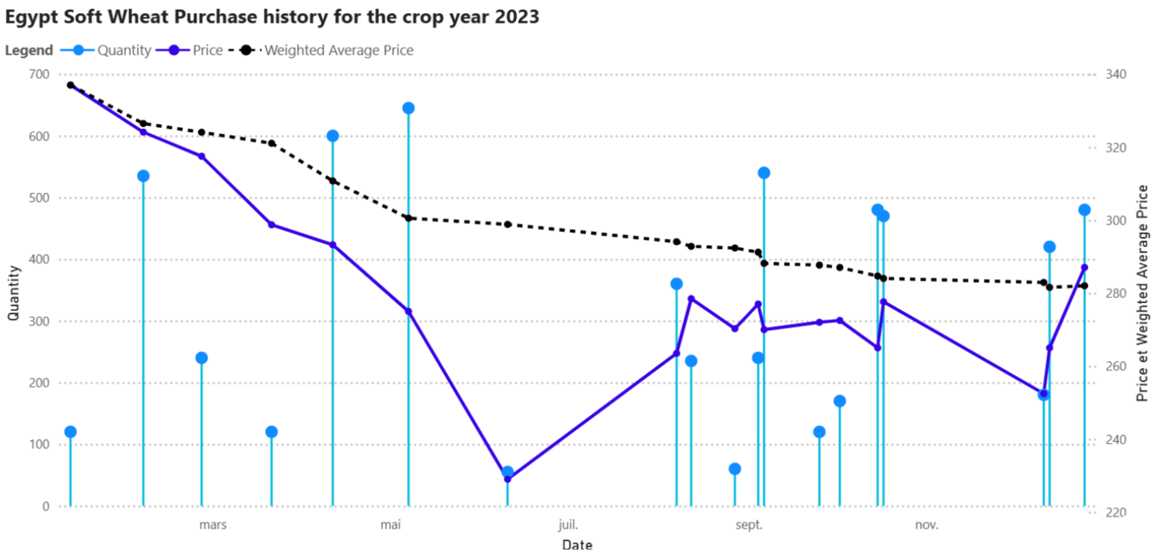


Figure 5.6: Egypt's Soft Wheat Purchase History for the 2023 Crop Year

Figure 5.6 presents Egypt's soft wheat purchase history for the crop year. We will analyze Egypt's purchasing strategy and assess whether their forecasting helped them save money. The blue line representing prices shows that Egypt consistently purchases at lower prices, as the prices are below the weighted average. However, let's examine the purchasing patterns in May 2023 and June 2023 more closely

Month	Quantity(MT)	Price($)	Total Price($)
May	645,000	247.98	159,947,100
June	55,000	229.00	12,595,000

Table 5.2: Egypt's Soft Wheat Purchases in May and June 2023

When analyzing Table 5.2. We notice that Egypt bought a significant quantity when prices were high in May. In contrast, in June, when prices decreased, only 55,000 MT were purchased. This is considered a poor decision resulting from inaccurate forecasting. If the quantities had been reversed, Egypt could have saved 11 million dollars.

5.6.2 Analysis of Tunisia's 2024 Purchasing Strategy

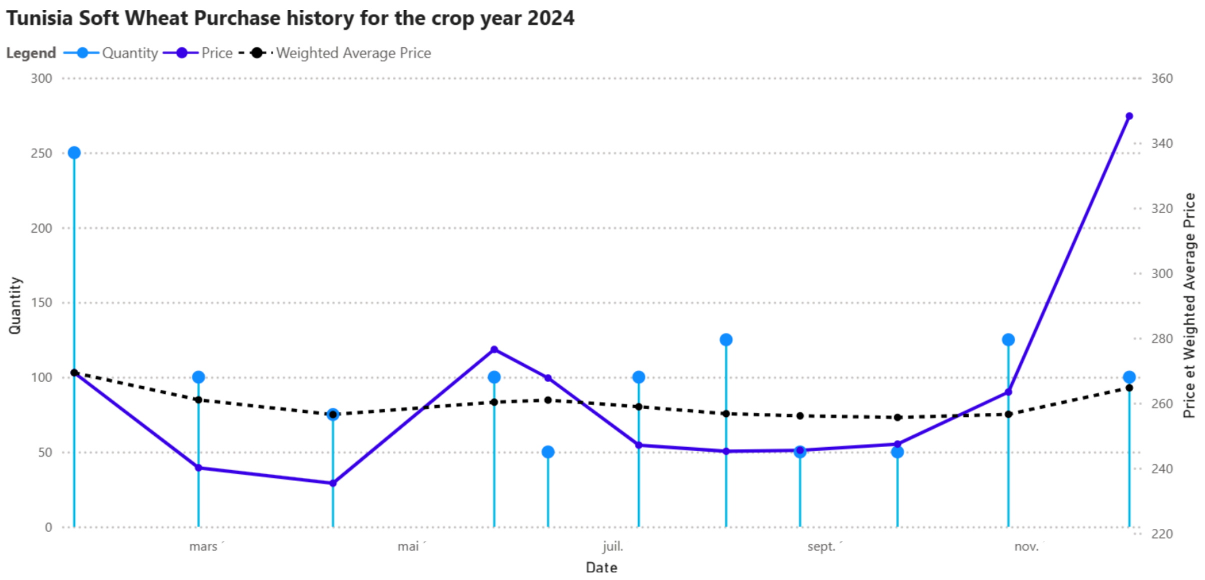


Figure 5.7: Tunisia's Soft Wheat Purchase History for the 2024 Crop Year

Figure 5.7 presents Tunisia's soft wheat purchase history for crop year 2024: When analyzing it we observe a strategic procurement approach. In most cases, the blue line representing the purchase price is below the weighted average price, indicating that Tunisia often secures wheat at a lower cost. This suggests a deliberate strategy of buying in larger quantities when market prices are relatively low, allowing the country to optimize costs and efficiently build reserves.

5.7 Conclusion

This chapter demonstrated the successful deployment of our automated tender processing system using Docker containerization and Apache Airflow orchestration. The comprehensive Power BI dashboards provide stakeholders with interactive visualizations to analyze tender data, price trends, and procurement patterns. The deployed solution ensures reliable, scalable operations while delivering actionable insights through user-friendly interfaces that support informed decision-making in agricultural commodity markets.

CONCLUSION AND PERSPECTIVES

This project is part of the requirements for obtaining an engineering diploma in Data Engineering and Decision Systems from the National School of Electronics and Telecommunications of Sfax. The internship was conducted at DNEXT, a leading international company specializing in agricultural commodity market intelligence. DNEXT's mission is to provide businesses with accurate and timely market data to drive strategic decisions. The company focuses on automating data collection and processing workflows to enhance efficiency and reduce human error.

This project addresses the challenges of manual tender data collection and processing by implementing an automated solution that streamlines the entire workflow. By integrating artificial intelligence techniques for data extraction and automated checks for data quality, the solution significantly reduces the time and effort required for manual tasks, while minimizing human error. The automated system ensures real-time, accurate updates to market intelligence dashboards, providing clients with reliable insights and enhancing operational efficiency.

Efforts will focus on enhancing the existing scripts to improve performance, handle more complex data scenarios, and further streamline the workflow, ensuring the system remains scalable, efficient, and adaptable to evolving requirements. By continuously improving and expanding the system, we can better meet the dynamic needs of the industry and provide clients with even more timely and accurate market intelligence.

WEBOGRAPHY

[1] **DNEXT Logo**. Official Website **[Online]**. Consulted on 16/05/2025. Available at: https://www.dnext.io/

[2] **Delivery terms**. What are Shipping Terms: Meaning And Glossary**[Online]**. Consulted on 18/05/2025. Available at: https://www.airsupplycn.com/shipping-term/

[3] **Scrum Life Cycle Logo**. What is Scrum Process Lifecycle? **[Online]**. Consulted on 18/05/2025. Available at: https://hygger.io/blog/what-is-scrum-lifecycle/

[4] **Langchain**. Introduction **[Online]**. Consulted on 24/05/2025. Available at: https://python.langchain.com/docs/introduction/

[5] **Apache Airflow**. What is Airflow®? **[Online]**. Consulted on 15/05/2025. Available at: https://airflow.apache.org/docs/apache-airflow/stable/index.html

[6] **Power BI**.What is Power BI? **[Online]**. Consulted on 20/05/2025. Available at: https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

[7] **Mongo DB**. what-is-mongodb **[Online]**. Consulted on 20/05/2025. Available at: https://www.mongodb.com/fr-fr/company/what-is-mongodb

[8] **Docker**. What is Docker? **[Online]**. Consulted on 21/05/2025. Available at: https://docs.docker.com/get-started/docker-overview/

[9] **Agrochart**. Website **[Online]**. Consulted on 21/05/2025. Available at: https://www.agrochart.com/news

[10] **ProspexAgro**. Telegram Channel **[Online]**. Consulted on 28/05/2025. Available at: https://web.telegram.org/k/@prospexagro

[11] **Containerizing**. A Step-by-Step Guide to Containerizing and Deploying Machine Learning Models with Docker! **[Online]**. Consulted on 24/05/2025. Available at: https://dev.to/pavanbelagatti/

Tenders Observatory:
News & Tenders Opportunities

Ghaith KRIFA

Abstract: Dnext aims to provide its clients with accurate, real-time market analysis in the trading of grains and oilseeds. This is achieved through the development of an advanced system that continuously feeds a dashboard with the necessary data using artificial intelligence. This project will minimize the time spent on data collection and quality control, while ensuring clients have constant access to the latest updates via the dashboard. Moreover, the system will ease the workload of the analyst team by automating these processes, thereby replacing manual tasks with faster and more efficient methods.

Key Words: Trading, Tenders, Grains, Oilseeds, Extracting Key Informations, Artifical Intelligence, Market Insights, Data Analysis, Reporting.

Résumé: Dnext vise à offrir à ses clients des analyses de marché précises et en temps réel dans le domaine du commerce des céréales et oléagineux. Cela passe par le développement d'un système avancé utilisant l'intelligence artificielle pour alimenter en continu un tableau de bord avec les données nécessaires. Ce projet permettra de minimiser le temps consacré à la collecte et au contrôle de la qualité des données, tout en assurant à nos clients un accès constant aux dernières mises à jour via ce tableau de bord. De plus, ce système facilitera le travail de l'équipe des analystes en automatisant ces processus, remplaçant ainsi les tâches manuelles par des méthodes plus rapides et efficaces

Mots Clés: Commerce, Appels d'offres, Céréales, Oléagineux, Extraction d'informations clés, Intelligence artificielle, Analyses de marché, Analyse de données, Reporting.

تلخيص: تهدف شركة Dnext إلى تقديم تحليلات دقيقة وآنية لعملائها في مجال تجارة الحبوب والزيوت النباتية. ويتم ذلك من خلال تطوير نظام متقدّم يعتمد على الذكاء الاصطناعي لتغذية لوحة قيادة بشكل مستمر بالبيانات اللازمة. سيساهم هذا المشروع في تقليص الوقت المخصص لجمع البيانات والتحقق من جودتها، مع ضمان وصول العملاء إلى أحدث التحديثات بشكل دائم عبر هذه اللوحة. علاوة على ذلك، سيسهّل هذا النظام عمل فريق المحللين من خلال الاستغناء عن العمليات اليدوية وتبنّي أساليب معالجة أكثر سرعة وفعالية.

الكلمات المفاتيح: التجارة، المناقصات، الحبوب، الزيوت النباتية، استخراج المعلومات الرئيسية، الذكاء الاصطناعي، تحليلات السوق، تحليل البيانات، التقارير.