THOMAS MORE

Lead Pipeline Traceability

Reflection

Ravin Shalmashi
Student Bachelor Applied Computer Science

2025 – 2026

Table of contents

1. Introduction

This reflection document discusses my experiences and insights gained during my internship at KBC Bank & Insurance within the Leads Capabilities (LeCa) team.

The internship focused on investigating data traceability issues within the LEDAS marketing pipelines. The objective was to understand why discrepancies existed between different pipeline stages and why some leads were not properly represented in the Data Traceability (DT) dataset.

This document consists of two main parts. The first part provides a substantive reflection on the project, including the achieved results, business impact, and recommendations for future improvements. The second part focuses on my personal development throughout the internship, including the challenges I faced, the skills I developed, and the lessons I learned.

2. Substantive reflection

The main objective of my internship was to investigate discrepancies within the LEDAS marketing pipelines and improve the understanding of lead traceability throughout the different processing stages.

At the beginning of the internship, the problem appeared relatively straightforward: certain leads were missing from the Data Traceability dataset, making it difficult to explain why some clients did not receive communications. However, as the investigation progressed, it became clear that the issue was much more complex than initially expected.

To analyse the problem, I first spent several weeks understanding the architecture of the LEDAS pipelines, the filtering logic, and the different stages involved in lead processing. I developed a methodology based on data reconciliation, where row counts and business keys were compared across multiple pipeline stages, including Offer Selection (OS), Offer Enrichment (OE), Lead Value (LV), and Data Traceability (DT).

One of the most important achievements of the project was identifying several root causes behind the observed discrepancies.

The first major finding was a duplication issue caused by an incorrect GDPR-related parent-child join. This issue generated multiple records for the same business entity and significantly inflated row counts within Offer Enrichment. After identifying the cause, the issue was fixed.

A second important finding was that buffer-related filtering was not included in the Data Traceability dataset. Initially, these leads appeared to be missing, but further investigation revealed that they were simply not represented in the traceability reporting. This finding explained a large portion of the remaining discrepancies.

Another issue was identified in the blocking table logic. The current implementation did not always select the most recent blocking record, which resulted in small but recurring inconsistencies. A solution was proposed by modifying the windowing logic to retain only the latest blocking end date.

Later in the internship, I performed a deeper investigation into the buffer update pipeline by executing the process locally and analysing intermediate outputs. This led to the identification of another duplication issue within the select_relevant_enrichments transformation. The root cause was traced back to dropping important uniqueness-defining columns, causing different records to become identical and resulting in duplicate generation.

Besides pipeline-related investigations, I also worked on improving the Qlik dashboard used for monitoring traceability. During this work, it became clear that several important metrics, such as buffer traceability values, were not available in the dashboard. I proposed multiple improvements, including the addition of reconciliation metrics and traceability-related indicators. These proposals were presented to the team and approved for further development.

The project is not entirely finished. While the major discrepancies have been explained and several issues have been fixed or documented, additional work can still be performed. For example, the recently discovered Fixed Selection Retrieval issue involving PTY_GUID_NO and HalPartyPartyRelation requires further investigation. Additionally, the proposed Qlik dashboard improvements still need to be fully implemented.

The results of this project provide significant value to KBC. The findings improve transparency within the lead processing pipelines, reduce the time required for debugging future issues, and increase confidence in reporting and monitoring. Furthermore, the debugging methodology and documentation created during the internship can be reused for future investigations.

For the future, I would recommend continuing the integration of traceability information into monitoring dashboards, implementing automated validation checks between pipeline stages, and maintaining detailed documentation of filtering logic and pipeline transformations.

3. Personal reflection

This internship has been one of the most valuable learning experiences of my academic career so far.

When I started at KBC, I quickly realized that I was entering a highly complex environment. The LEDAS pipelines process large amounts of data every day, involve many interconnected components, and contain years of accumulated business logic. At first, understanding how everything worked felt overwhelming. There were many systems, tools, and codebases that I had never worked with before, and I often felt like I had more questions than answers.

One of the biggest challenges I faced was learning how to navigate such a large and complex system. During my first weeks, I even spent time studying the wrong pipeline files because I misunderstood part of the architecture. Instead of becoming discouraged, I used this experience as a learning opportunity. I started asking more questions, scheduling meetings with team members, validating my assumptions, and documenting everything I learned. This significantly improved both my understanding and my confidence.

Throughout the internship, I developed several technical skills. I improved my ability to analyze large data pipelines, work with PySpark, understand data transformations, investigate discrepancies, and perform root cause analysis. I also gained practical experience with Airflow, Qlik, Bitbucket, and large-scale data engineering workflows.

In addition to technical skills, I developed important professional competencies. Communication became a key part of my daily work. I regularly presented findings to team members, discussed technical issues with experienced engineers, documented investigations, and translated complex technical concepts into understandable explanations.

Another important lesson was learning how to approach problems systematically. Initially, I often focused directly on the symptoms of an issue. Over time, I learned to first build a clear understanding of the overall system, define a structured methodology, and then narrow down the problem step by step. This approach proved essential when investigating complex discrepancies involving multiple root causes.

I also learned the importance of persistence. Many of the issues I investigated required days of analysis before meaningful progress was made. There were moments when I felt stuck, especially when dealing with access limitations, unexpected results, or highly distributed logic. However, these situations taught me that solving complex problems often requires patience, experimentation, and continuous validation.

Looking back, I believe I have grown significantly both technically and professionally. I am more confident when approaching unfamiliar systems, more comfortable communicating with stakeholders, and more capable of conducting structured investigations in complex technical environments.

Overall, this internship has given me valuable real-world experience and has strengthened my interest in data engineering and analytics. The combination of technical challenges, teamwork, problem-solving, and continuous learning made this internship a highly rewarding experience that will have a lasting impact on my future career.