Project Plan

Lead Pipeline Traceability

Internship – Applied Computer Science

Thomas More University of Applied Sciences

1. Introduction

This document presents the project plan for the internship conducted at KBC Bank & Insurance as part of the Applied Computer Science program at Thomas More University of Applied Sciences. The internship takes place over a 14-week period starting on February 23rd.

The project focuses on investigating inconsistencies in the lead processing pipeline used for marketing communications within KBC. Specifically, the project aims to identify why certain leads and clients are missing from the Data Traceability file, which should record all filtered leads and their corresponding filter-out reasons.

The goal of this document is to describe the background, objectives, scope, planning, and reporting structure of the internship project.

2. Company Background

2.1 KBC Bank & Insurance

KBC is a Belgian financial institution providing integrated banking and insurance services to individuals and businesses. The organization relies heavily on data driven systems to support marketing campaigns, customer communication, and digital banking services.

To manage marketing communication with clients, KBC uses a lead processing pipeline that determines which customers receive certain communications or offers through various channels such as email, mobile notifications, or direct messaging.

2.2 Lead Capabilities (LeCa)

During the internship, the work is conducted within the LeCa (Lead Capabilities) group, which is responsible for managing and improving the lead generation and processing systems used in marketing campaigns.

The LeCa team maintains data pipelines that:

- generate potential leads
- process and enrich lead information
- apply filtering rules
- determine which clients receive communications

These pipelines process large volumes of data and rely on several data engineering technologies to ensure that communications are targeted correctly.

3. Current Situation

The lead processing system consists of multiple data pipelines that process leads through different filtering stages. These stages apply business rules and technical constraints to determine which leads are valid for communication.

During this process, some leads are filtered out due to conditions such as:

- duplicate messages
- missing consent (GDPR)
- contactability restrictions
- commercial approval rules
- language mismatches
- channel limitations
- journey prioritization rules

Whenever a lead is filtered out, the system should record the exclusion in a Data Traceability file, including the reason why the lead was removed.

However, inconsistencies have been identified between:

- the total number of leads processed
- the number of filtered leads recorded in the Data Traceability file

This indicates that some leads are being removed from the pipeline without being properly logged with a reason.

4. Project Objective

The main objective of this project is to investigate the missing lead and client entries in the Data Traceability file and ensure that all filtering events in the pipeline are properly recorded.

The project will focus on:

- identifying where leads disappear in the pipeline
- analysing filtering stages that may not log exclusion reasons
- determining the technical cause of missing traceability entries
- proposing or implementing fixes where possible
- validating that the pipeline records filtered leads correctly
- documenting the findings and improvements

By the end of the internship, the system should provide complete and reliable traceability for filtered leads, improving the transparency and maintainability of the pipeline.

5. Business Value

This project provides several benefits for KBC.

Improved Data Transparency

Ensuring that all filtered leads are recorded with a reason allows teams to understand why certain clients did not receive communications.

Better Monitoring

The Data Traceability file is used in dashboards and monitoring tools. Missing information can lead to incorrect reporting and misinterpretation of campaign performance.

Faster Troubleshooting

When pipeline issues occur, engineers need to investigate the source of the problem. Proper traceability significantly reduces the time required to debug pipeline behaviour.

Compliance and Governance

Many filtering rules are related to privacy regulations such as GDPR consent and contactability rules. Accurate logging ensures compliance with internal policies and legal requirements.

6. Technical Environment

The lead processing pipelines rely on a modern data engineering stack.

The following tools and technologies are used during the project:

Data Processing

- PySpark for large-scale data processing
- Glue pipelines for pipeline orchestration and data processing

Development Environment

- Jupyterhub for exploring and analysing pipeline data
- Bitbucket for source code management

Workflow Management

- Airflow for scheduling and monitoring pipeline executions

Analytics and Monitoring

- Qlik dashboards for monitoring pipeline results and analysing discrepancies

Marketing Integration

- Adobe Campaign for managing marketing communications and campaign configurations

These technologies together form the ecosystem in which the lead pipelines operate.

7. Project Scope

7.1 In Scope

The project focuses on analysing and improving the traceability of leads in the pipeline.

The following tasks are included:

- Understanding the architecture of the lead processing pipeline • Identifying filtering stages where leads may disappear
- Analysing pipeline behaviour through logs, dashboards, and code • Identifying the root causes of missing traceability entries • Proposing and implementing technical fixes where possible • Testing the corrected pipeline behaviour
- Improving the dashboard
- Documenting the filtering logic and improvements

7.2 Out of Scope

The following tasks are not part of the project:

- redesigning the entire lead pipeline architecture
- creating new marketing campaigns
- modifying business targeting strategies
- developing new marketing tools

The focus remains strictly on traceability and pipeline investigation.

8. Project Planning

The internship lasts 14 weeks, during which the project will progress through several phases.

Phase 1 – Onboarding and System Understanding

Weeks 1–2

- Introduction to the LeCa team and working methodology
- Understanding the overall LEDAS architecture
- Reviewing internal documentation
- Exploring Bitbucket repositories
- Learning the purpose of the different pipeline layers
- Gaining access to Airflow, JupyterHub, Qlik, and source repositories
- Initial analysis of Level 5 and LEDAS pipelines

Phase 2 – Pipeline Exploration

Weeks 3-4

- Deep analysis of Level 5 and LEDAS pipelines
- Mapping the complete lead flow from Fixed Selection Retrieval to Data Traceability
- Studying filtering logic and exclusion mechanisms
- Developing the first analysis notebooks
- Evaluating different approaches for tracing discrepancies

Phase 3 – Root Cause Investigation

Weeks 5–8

- • Developed reusable PySpark notebooks for discrepancy analysis
- Compared counts between:
    - Offer Selection (OS)
    - Offer Enrichment (OE)
    - Lead Value (LV)
    - Data Traceability (DT)
- Investigated multiple marketing solutions
- Validated discrepancies across datasets
- Performed root cause analysis on unexplained differences

Phase 4 – Buffer Pipeline Investigation

Weeks 9–10

- Executed buffer update locally
- Traced intermediate outputs step-by-step
- Compared transformation results
- Investigated duplicate generation within buffer processing\

Phase 5 – Documentation

Weeks 11-12

- Creation of internal Confluence documentation
- Documentation of:
    - methodology
    - debugging approach
    - root causes
    - findings
    - recommendations
- Creation of university realization document
- Presentation of findings to the LeCa team

Phase 6– Qlik Dashboard Improvement

Weeks 11-13

- improving existing visualizations
- validating dashboard calculations
- adding new indicators if necessary
- ensuring filtered leads and filter reasons can be easily compared

Phase 7 – Final Report

Week 14

- Finalize realization document
- Finalize Confluence documentation
- Complete Qlik improvement proposal
- Present findings and recommendations
- Handover investigation methodology and documentation

9. Progress Reporting

Regular communication is maintained throughout the

project. **Weekly Status Reports**

A weekly report is submitted describing:

- tasks completed during the week
- current progress
- encountered challenges
- planned tasks for the next week

Team Meetings

The internship includes regular meetings with the LeCa team.

These include:

- Two stand-up meetings per week
- Weakly meetings with team members to check on progress
- additional meetings when required
- discussions with the mentor regarding project progress

11. Expected Deliverables

By the end of the internship, the following deliverables are expected:

- Analysis of missing leads in the Data Traceability file
- Identification of the root causes of missing traceability records
- Proposed and implemented improvements to the pipeline
- Validation results demonstrating corrected behaviour
- Improved dashboard
- Final internship report documenting the project.