2024

Data Observability

PROJECT PLAN REPORT

WALID BIROUK

TOGAETHER | ALGORHYTHM

Introduction

This document outlines the comprehensive efforts and outcomes of my four-month internship at Algorhythm, a premier data consultancy firm based in Belgium. The primary objective of this internship was to conduct a comparative study of Data Observability tools available in the market, evaluate their capabilities, and determine the best fit for Algorhythm's managed services offering. The scope of the project evolved to include the development of a Dockerized modern data stack pipeline and the implementation of a Proof of Concept (POC) using OpenMetadata. This report is divided into two main sections: a substantive reflection on the project activities and accomplishments, and a personal reflection on the professional and personal growth experienced during the internship. The goal is to provide a clear and detailed account of the project’s execution, its impact on Algorhythm, and the learning outcomes derived from this internship experience.

1 Introduction to Algorhythm

Algorhythm is a leading data consultancy firm based in Belgium, specializing in data strategy, data science, data architecture, data engineering, and data governance. They offer services such as extending staff with data experts, delivering end-to-end projects, and setting up data infrastructure. The firm is proficient in cloud platforms (Azure, AWS, Google, Oracle, Snowflake) and front-end tools (Power BI, Tableau, Qlik).

Algorhythm provides tailored solutions for various data needs, from data strategy to machine learning for decision support or automation.

2 Assignment Description

The purpose of this internship assignment is to conduct a comparative study of available Data Observability tools on the market, evaluate their capabilities, and determine which tool best fits within our managed services offering. This task involves an in-depth market analysis to evaluate and compare the tools.

2.1 Project Evolution

Beyond the initial scope of the comparative study, the project evolved to include the development of a Dockerized modern data stack pipeline and the implementation of a Proof of Concept (POC) using OpenMetadata, tested against predefined selection criteria. These additional steps were undertaken by me, to enhance the study and provide practical insights into the implementation and effectiveness of the chosen tool.

2.2 Problem Statement

With the ever-increasing demand for data, enterprises face significant challenges in managing complex data pipelines. Key issues include data quality problems, infrastructure monitoring difficulties, and budgetary constraints. The objective of this project is to address these challenges by evaluating and selecting a data observability tool that ensures data quality, reliability, and comprehensive monitoring across diverse data sources. This tool will be integrated into Algorhythm's managed services offering, enhancing our capability to deliver robust data governance and observability solutions.

3 Detailed Project Approach and Timeline

This section outlines the step-by-step approach taken to complete the project, along with the timeline for each phase. The approach is designed to ensure a thorough evaluation of data observability tools, successful implementation of the POC, and comprehensive analysis and documentation of the findings.

3.1 Visual Timeline (Gantt Chart)

The Gantt chart provided visually represents the timeline for each project task.

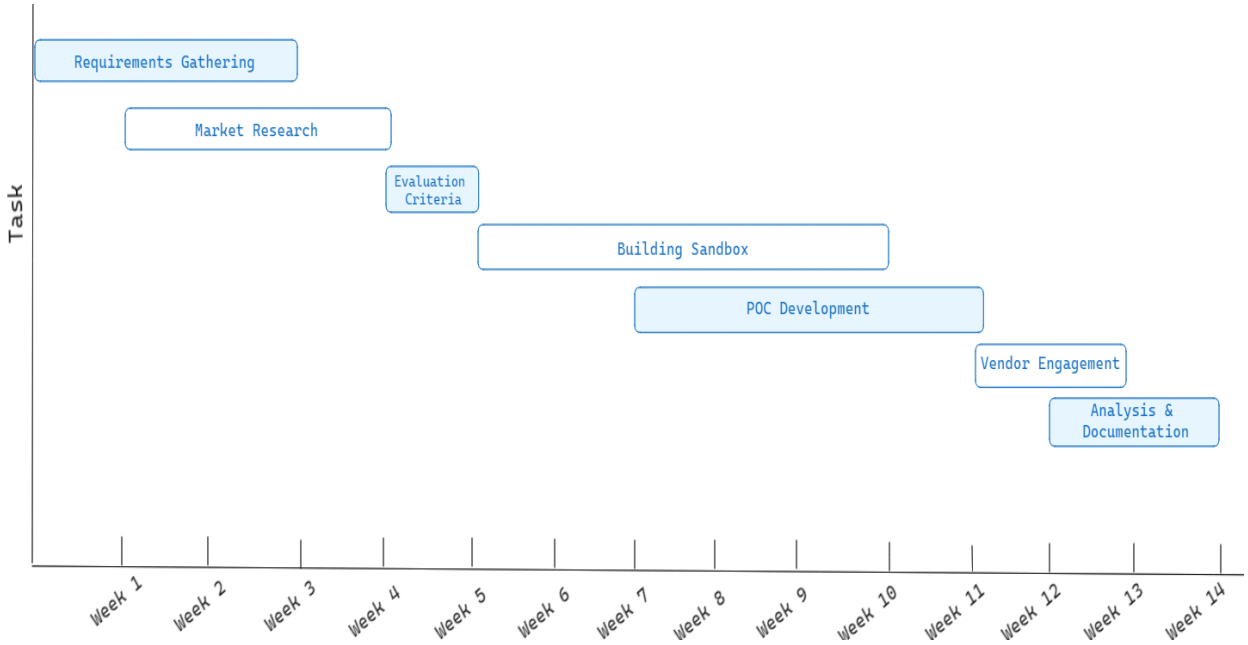

Figure 1: Gantt Chart for the Data Observability Comparative Study and POC Implementation Project

3.2 Week 1 - 3: Objectives and Requirements Gathering

- **Objective**: Understand the specific needs and pain points related to data observability within Algorhythm and its clients.
- **Tasks**:
  - Conduct meetings with key stakeholders.
  - Document detailed requirements.
  - Identify objectives for successful data observability.

3.3 Week 2 - 4: Market Research

- **Objective**: Identify and analyse available data observability tools on the market.
- **Tasks**:
  - Conduct a literature review on data observability.
  - Analyse features, strengths, and weaknesses of various tools.
  - Shortlist tools for further evaluation

3.4 Week 4 - 5: Evaluation Criteria Development

- **Objective**: Develop criteria for evaluating the shortlisted tools.
- **Tasks**:
    - Define metrics and architecture for assessing tool capabilities (e.g., integration, scalability, cost).
    - Create a classification table for tool comparison.
    - Validate criteria with stakeholders.

3.5 Week 5 - 10: Building the Sandbox Environment

- **Objective**: Set up a comprehensive sandbox environment for testing data observability tools.
- **Tasks**:
    - Configure Dockerized modern data stack.
    - Integrate tools such as Airflow, Airbyte, dbt, Spark, Hive Metastore, S3 and Postgres DWH.
    - Ensure the environment is robust and mirrors real-world data pipelines.

3.6 Week 7 - 11: Proof of Concept (POC) Implementation

- **Objective**: Implement and test the selected data observability tool (OpenMetadata) in the sandbox environment.
- **Tasks**:
    - Configure OpenMetadata and connect it to the data stack.
    - Run comprehensive tests to monitor data quality, freshness, schema, lineage, and volume.
    - Evaluate real-time alert capabilities and data governance features.

3.7 Week 11 - 13: Vendor Engagement

- **Objective**: Engage with vendors to understand pricing models and negotiate terms.
- **Tasks**:
    - Contact vendors of shortlisted tools.
    - Gather detailed pricing and feature information.
    - Compare costs and benefits to determine the best fit.

3.8 Week 12 - 14: Analysis and Documentation

- **Objective**: Document findings and provide recommendations.
- **Tasks**:
    - Analyse POC results against evaluation criteria.
    - Compile a comprehensive report detailing the study and findings.
    - Present recommendations to stakeholders.

Conclusion

The project successfully identified and implemented OpenMetadata as the preferred data observability tool for Algorhythm. Through a comprehensive evaluation process, including market research, detailed analysis, and practical testing, the project achieved its primary objective of enhancing data quality, monitoring, and governance within Algorhythm's managed services.

The Proof of Concept (POC) demonstrated OpenMetadata's effectiveness, meeting all predefined success criteria. This implementation not only validated the tool's capabilities but also provided valuable insights into its practical application in real-world scenarios. The development of a Dockerized modern data stack pipeline further enriched the project's scope, offering a robust testing environment that closely mirrors actual operational conditions.

Future steps involve scaling the solution by deploying it in Kubernetes, integrating it with additional systems such as Redshift and BigQuery, and enhancing security measures. Continuous training and support for staff will ensure the tool's effective use, while regular monitoring and adjustments will maintain optimal functionality.

Overall, this project has significantly strengthened Algorhythm's ability to deliver robust data governance and observability solutions, positioning the company for continued success in managing complex data environments.