Projeto de Formatura – 2023

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

Tema:

Ferramentas Computacionais para Gestão da Qualidade de Dados

Motivação

Em projetos de monitoramento ambientais que envolvem coleta de dados através de sensores por um longo período de tempo, a aquisição de dados está sujeita a problemas na operação dos equipamentos e no seu mal funcionamento. O problema de tratar a qualidade de dados é fundamental no processo e experimentos de Ciência dos Dados [1] o qual vem sendo abordado no Laboratório de Big Data da Escola Politénciica da USP, através da ferramenta DataMap/Amazon.

Dado o alto volume de dados tratados, é explícita a importância de se monitorar e administrar a entrada destes no sistema, de forma a garantir sua qualidade e posterior uso pela comunidade científica.

Como referência bem sucedida nessa área, pode-se citar o Programa do Departamento de Energia dos EUA, o portal do Atmospheric Radiation Measurement (ARM), que reúne dados de diversos centros de pesquisa ao redor do mundo. Contudo, esta é uma plataforma fechada e cujo uso é limitado a pesquisadores associados ao ARM, de forma que não-membros podem apenas acessar os dados disponibilizados, mas não inseri-los.

Objetivos

O objetivo deste trabalho é a criação de uma ferramenta para auxiliar os pesquisadores em todas as etapas de seu Workflow e garantir que os dados coletados estejam de acordo com os princípios de boa governança.

Arquitetura

A arquitetura desenvolvida é monolítica, e, portanto, todas as funcionalidades estão contidas em uma única base de código. Ela foi elaborada com base no fluxo de interações esperado do usuário, que pode ser vista no diagrama a seguir, em formato BPMN (Business Process Model Notation).

Todos os dados relevantes são salvos em um cluster MongoDB, um banco de dados NoSQL, especialmente relevante por não limitar o usuário a um único esquema de inserção possível.

A comunicação com o MongoDB é feita diretamente no front-end, a partir do uso do Streamlit, um framework em python, que atua em conjunto com a biblioteca Pandas e Beanie também para as etapas de visualização de dados e demais funções do backend necessárias

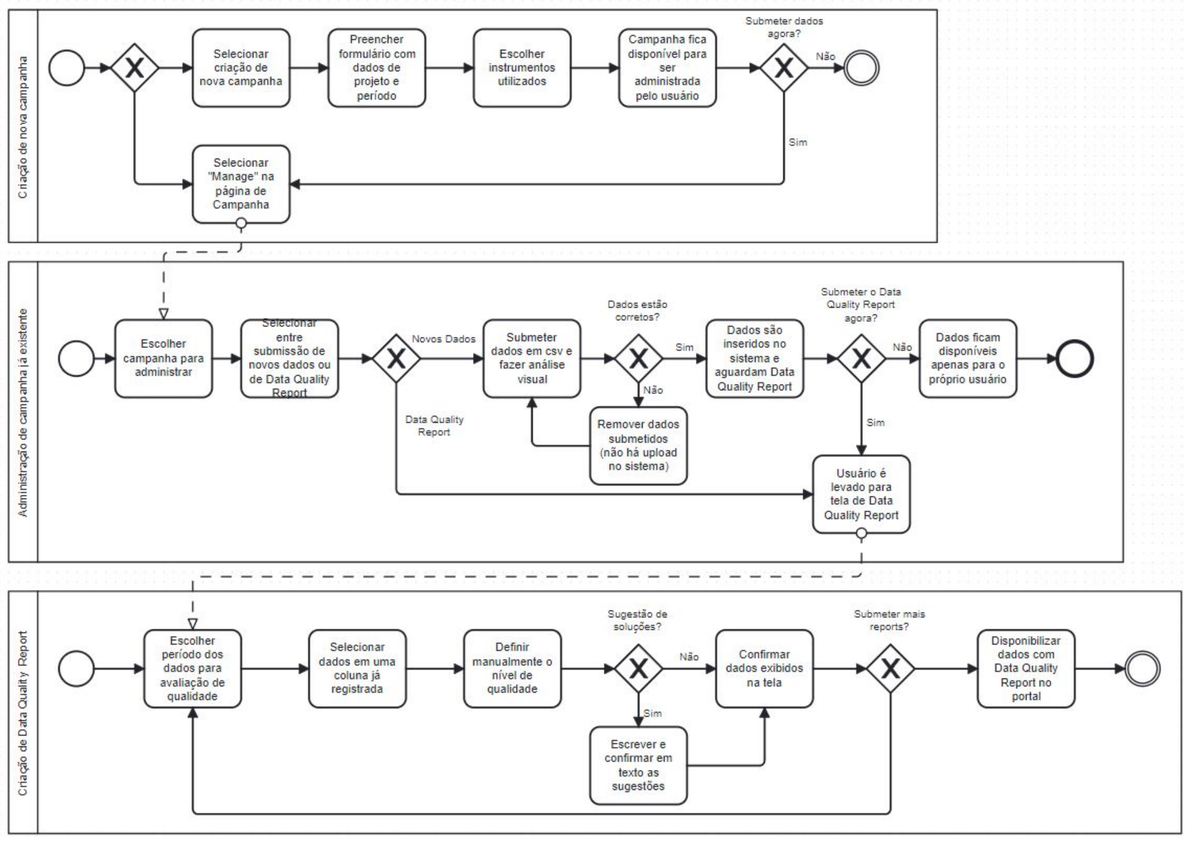

Figura 1: Diagrama BPMN de interações do usuário com a plataforma

Funcionalidades e resultados

O desenvolvimento do POC (Proof of Concept) com as ferramentas planejadas implicou na especificação dos requisitos funcionais, e capacidades mínimas esperadas da plataforma:

I. Criação de novas Campanhas: O usuário cadastrado pode criar uma campanha com período definido para organizar seus dados
II. Inserção e recuperação de dados: O usuário deve ser capaz de inserir e manipular seus dados dentro da plataforma.
III. Geração de Data Quality Reports: Para cada fluxo de dados individual inserido, o usuário deve ser capaz de gerar um Data Quality Report associado a ele.
IV. Disponibilização para download: Após geração do Data Quality Report, os dados devem ficar disponíveis para todos os usuários (inclusive não cadastrados)

Nesse sentido, o produto foi desenvolvido com sucesso e todas as funcionalidades acima foram devidamente implementadas. As figuras abaixo contém algumas das interfaces elaboradas.

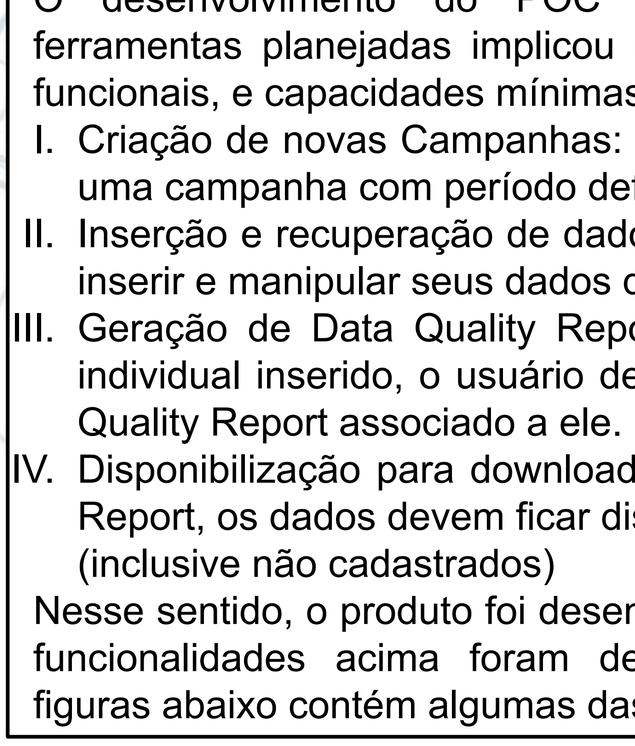

Figura 2: Página de criação de novas campanhas

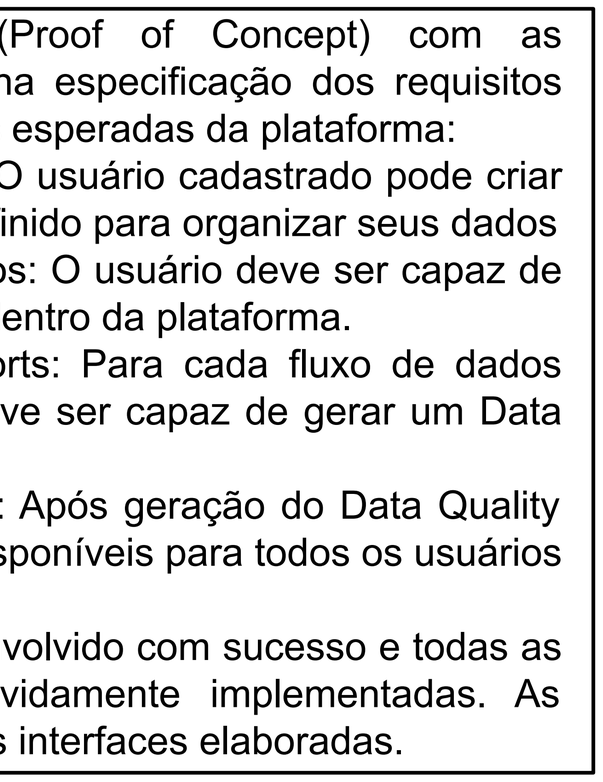

Figura 3: Página de geração de Data Quality Report

Integrantes: Gabriel Gandra Prata Gonçalves

Professor(a) Orientador(a): Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Co-orientador(a): Felipe Valencia Almeida