Projeto de Formatura – 2023 – Press Release
PCS - Departamento de Engenharia de Computação e Sistemas Digitais
Engenharia Elétrica – Ênfase Computação

Tema:

Ferramentas Computacionais para Gestão da Qualidade de Dados

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 mau funcionamento. O problema de tratar a qualidade de dados é fundamental no processo e experimentos de Ciência dos Dados o qual vem sendo abordado no Laboratório de Big Data da Escola Politécnica 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.

O objetivo deste trabalho foi 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. Dessa forma, era esperado que fosse projetada e implementada uma interface simplificada para que se possa realizar a inserção de dados de forma padronizada, com capacidades relevantes de controle da qualidade de dados; e também que esteja disponível um banco de dados onde estes ficarão armazenados.

O desenvolvimento do POC (*Proof of Concept*) com as ferramentas planejadas foi feito em arquitetura monolítica, utilizando *MongoDB* (banco de dados NoSQL) e *Streamlit* (framework para Python) implicou na especificação dos requisitos funcionais, e definição capacidades mínimas esperadas da plataforma:

1. Criação de novas Campanhas: O usuário cadastrado pode criar uma campanha com período definido para organizar seus dados.
2. Inserção e recuperação de dados: O usuário deve ser capaz de inserir e manipular seus dados dentro da plataforma.
3. 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.
4. 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.

Integrantes: Gabriel Gandra Prata Gonçalves

Professor(a) Orientador(a): Prof. Dr. Pedro Luiz Pizzigatti Corrêa
Co-orientador(a): Felipe Valencia Almeida