GABRIEL GANDRA PRATA GONÇALVES

FERRAMENTAS COMPUTACIONAIS PARA GESTÃO DA QUALIDADE DE DADOS

São Paulo
2023

GABRIEL GANDRA PRATA GONÇALVES

FERRAMENTAS COMPUTACIONAIS PARA GESTÃO DA QUALIDADE DE DADOS

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro Eletricista com ênfase em Computação.

São Paulo
2023

GABRIEL GANDRA PRATA GONÇALVES

FERRAMENTAS COMPUTACIONAIS PARA GESTÃO DA QUALIDADE DE DADOS

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro Eletricista com ênfase em Computação.

Orientador:

Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Co-orientador:

Felipe Valencia de Almeida

São Paulo
2023

RESUMO

Tanto em projetos científicos como em projetos corporativos é necessário que os dados utilizados tenham a qualidade adequada para que possam ser utilizados em experimentos em Ciência dos Dados. A ausência de ferramentas de uso livre e aberto, especialmente em projetos científicos, é um dos obstáculos para que resultados de pesquisas e analises sejam realmente usados para gerar valor e conhecimento. A partir desses desafios, o principal objetivo deste trabalho é criar uma plataforma abrangente, inspirada pelo portal da Atmospheric Radiation Measurement (ARM), para auxiliar os pesquisadores em todo o seu fluxo de tratamento de dados, garantindo a governança de qualidade dos dados a partir de *Data Quality Reports*. A plataforma produzida a partir das ferramentas projetadas foi inicialmente implementada com foco em pesquisa na Amazônia a partir da colaboração com o AmazonFACE, e visa oferecer uma interface padronizada para inserção de dados, capacidades de controle de qualidade, um banco de dados para armazenamento e acesso livre para partes interessadas. O desenvolvimento da arquitetura foi baseado em técnicas de projeto abrangentes, buscando integrar visões diferentes para facilitar tanto o entendimento de possíveis contribuidores quanto a implementação prática. Este projeto evidenciou tanto as vantagens de um desenvolvimento aberto focado em qualidade de dados, quanto seus desafios e oportunidades para aprimoramentos futuros.

Palavras-Chave – Qualidade de dados, Plataforma Digital, Amazônia.

ABSTRACT

In both scientific and corporate projects, it is necessary for the data used to have the appropriate quality to be employed in Data Science experiments. The absence of free and open-source tools, especially in scientific projects, is one of the obstacles for research and analysis results to be effectively used to generate value and knowledge. From these challenges, the main objective of this work is to create a comprehensive platform, inspired by the Atmospheric Radiation Measurement (ARM) portal, to assist researchers throughout their data processing workflow, ensuring data quality governance through Data Quality Reports. The platform, produced from the designed tools, was initially implemented with a focus on research in the Amazon through collaboration with AmazonFACE. It aims to provide a standardized interface for data input, quality control capabilities, a database for storage, and open access for stakeholders. The architecture development was based on comprehensive design techniques, aiming to integrate different perspectives to facilitate both the understanding of potential contributors and practical implementation. This project highlighted the advantages of open development focused on data quality, as well as its challenges and opportunities for future enhancements.

Keywords – Data Quality, Digital Platform, Amazon Rainforest.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

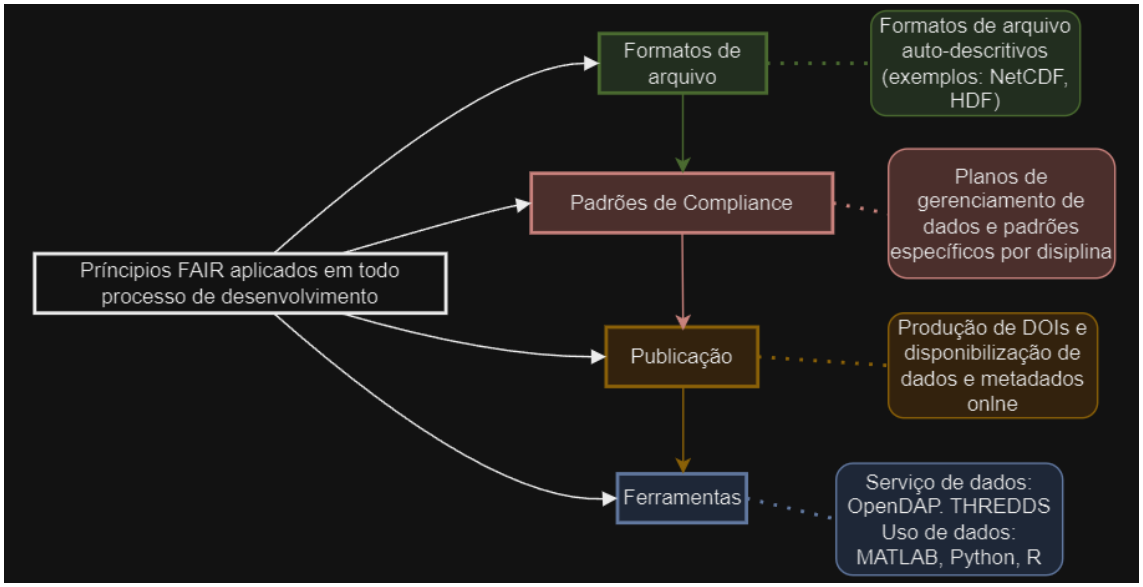
Há entre pesquisadores colaboradores do autor deste trabalho um esforço para o desenvolvimento de uma ferramenta denominada *DataMap*, que será de uso público e irá agregar tanto *DataViz* (Visualização de Dados) em mapas, séries temporais, *DataCubes*, dentre outros; quanto modelos preditivos e os próprios *DataSets* (conjuntos de dados) brutos e tratados que foram e serão coletados. Espera-se também que ele possa ser usado e modificado livremente, no formato *Open Source*.

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.

Nota-se que uma grande inspiração para o projeto é o portal do *Atmospheric Radiation Measurement* (ARM), organização americana para pesquisas atmosféricas que reúne dados de diversos centros de pesquisa ao redor do mundo.

Neste contexto, dois padrões importantes, que serão usados como base, são os definidos pelo próprio ARM [PEPPLER et al., 2016], que estabelecem maneiras relevantes de monitoramento de qualidade no contexto especifico de pesquisas atmosféricas; e os princípios FAIR *(Findability, Accessibility, Interoperability, and Reusability)*, mais abrangentes e que buscam garantir que a pesquisa produzida a partir dos dados obtidos possa ser sempre reprodutível, por meio do reuso de dados. Na figura 1 podemos ver uma das possíveis formas de elaborar uma aplicação dentro deles.

Figura 1 – Exemplo de como construir aplicações que obedecem aos princípios FAIR



Fonte: Elaboração própria.

Em um primeiro momento, as ferramentas desenvolvidas serão testadas no projeto AmazonFACE, que busca estudar os impactos causados pelas mudanças ambientais na flora, fauna, e nos padrões de captura de carbono da região da floresta Amazônica. Sendo assim, os principais dados utilizados serão dados coletados pelas torres instaladas. Também é possível que sejam acrescidos na plataforma imagens e dados meteorológicos de satélites, bem como dados coletados de torres de monitoramento no solo e de aeronaves especializadas, produzidos por instituições como o INMET, o IBGE, a NASA, dentre outros.

Espera-se que a partir do resultado bem sucedido da implementação da plataforma neste contexto seja possível generalizá-la para integração em projetos mais amplos.

1.2 Objetivo

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. Desta forma, espera-se ao final do projeto que estejam disponíveis para os usuários uma interface simplificada para que se possa realizar a inserção de dados de forma padronizada, com capacidades relevantes de controle da qualidade de dados; que esteja disponível um banco de dados onde estes ficarão armazenados, e que outros pesquisadores também possam ter acesso a todas as informações submetidas, para que possam usá-las em suas próprias pesquisas.

1.3 Justificativa

O acesso aberto e facilitado a dados de boa qualidade é um desafio em muitos domínios e campos de conhecimento, não se limitando apenas aos fins acadêmicos, mas também aos jornalísticos, comerciais, legais, dentre outros. Neste campo, diversas ferramentas já foram criadas ao longo dos anos para auxiliar nesse processo, com níveis variados de sucesso em seus objetivos. Embora ferramentas de busca já sejam amplamente acessíveis e utilizadas por uma grande parcela da população, é notável como a confiabilidade dos resultados encontrados muitas vezes não pode ser garantida mesmo por organizações com recursos consideráveis. Este problema é particularmente relevante para a comunidade científica, que depende fundamentalmente de dados altamente confiáveis para extrair resultados validáveis [HAZEN et al., 2014] e [BATINI et al., 2009].

Neste contexto, a criação de uma plataforma unificada, com capacidade de monitoramento de qualidade associada a pesquisadores qualificados, é um avanço que permite que estes tenham maior acesso aos datasets relevantes para suas pesquisas. Tratando-se especificamente de dados ambientais, em seu escopo inicial, também será possível abordar um campo particularmente importante, especialmente considerando-se discussões atuais sobre mudanças climáticas. Acredita-se que a proposta discutida a seguir será capaz de suprir essas necessidades.

Em adição a isso, é perceptível que ferramentas disponíveis atualmente não estão, em sua maioria, disponibilizadas de forma livre em formato *Open Source*. Essa situação dificulta seu aprimoramento e atrasa o progresso científico, na medida em que ofusca o fluxo dos dados uma vez que estes são inseridos em cada ferramenta proprietária. Com desenvolvimento aberto, será possível que inúmeros pesquisadores tenham acesso não apenas aos dados, mas também as próprias ferramentas utilizadas para armazena-los e processá-los.

Por fim, nota-se que o projeto trará benefícios diretos para os pesquisadores do Amazon-FACE, que poderão contar com ferramentas robustas para garantir a qualidade dos dados trabalhados por eles, bem como a governança responsável destes. Espera-se que essa cooperação resulte em progresso

1.4 Organização do Trabalho

De maneira a expor a forma como este trabalho foi desenvolvido, há a divisão em seis capítulos ao todo, incluindo este. Este primeiro capítulo contém a introdução e as justificativas e motivações para realização do projeto. No segundo capítulo estão elencados os aspectos

conceituais relevantes nos quais a plataforma se baseia, apresentando definições de qualidade de dados, e trabalhos previamente publicados que influenciaram o desenvolvimento proposto. No terceiro capítulo é apresentada a metodologia utilizada, levando-se em conta a divisão do projeto nas ferramentas apresentadas, as etapas de desenvolvimento, os testes de validação e a implementação completa da plataforma. No quarto capítulo há a especificação de requisitos, a partir de discussões com os orientadores, pesquisadores do AmazonFACE, análises baseadas em ferramentas semelhantes e no cumprimento dos objetivos propostos. Já o quinto capítulo apresentará a implementação completa de todas as propostas anteriores, a serem desenvolvidas

Por fim, o último capítulo conterá as considerações finais, e irá apresentar as reflexões acerca das contribuições do trabalho, levando-se em conta quais objetivos foram efetivamente cumpridos e levantando as possibilidades de desenvolvimentos futuros a partir deste.

2 ASPECTOS CONCEITUAIS

Neste capitulo será exposta toda a fundamentação teórica utilizada como base para este projeto. Há alguns focos para esta validação: o primeiro são os conceitos relevantes de Qualidade de Dados, conforme definidos por fontes relevantes; em segundo são os conceitos extraídos da plataforma ARM, que inspirou o desenvolvimento do projeto. Por fim, também é tratado brevemente sobre o AmazonFACE, para o qual o projeto será inicialmente preparado como forma de validação e verificação de capacidades.

2.1 Qualidade de Dados

A Qualidade de Dados é um dos grandes focos deste projeto, sendo a sua garantia uma consideração em todos os pontos do desenvolvimento. Na literatura pode-se encontrar um conjunto amplo e diverso de definições para este termo, de forma que para se possa realizar um trabalho concreto faz-se necessário escolher uma definição mais adequada ao cenário escolhido. Neste caso, o ponto de partida escolhido foram os princípios FAIR, conforme definidos por [WILKINSON et al., 2016]. É importante ressaltar que, embora a adequação a estes princípios seja encorajada, dados os objetivos de garantir o uso da plataforma por pesquisadores independentes e de múltiplas origens, torna-se impossível garantir que todos os dados os cumpram integralmente. Dessa forma, o foco neste caso será em se certificar que estes dados estejam corretamente rotulados em seu nível de adequação, para que todos que tentem acessa-los estejam plenamente cientes de suas limitações

Os princípios FAIR são divididos nas quatro palavras do acrônimo: *Findable, Accessible, Interoperable, Reusable*. Cada uma destas palavras possui um subconjunto de itens associados, podendo-se assim ter um entendimento mais completo da qualidade dos dados analisados.

- *Findable* (Localizável)

  O primeiro passo é garantir que os dados e metadados sejam facilmente encontráveis por humanos e por computadores. Metadados em formatos que são legíveis por instrumentos

automatizados são essenciais para descoberta automática de datasets e serviços

  - F1 São atribuídos aos dados e metadados um identificador global único e persistente
  - F2 Dados são descritos com metadados ricos, conforme definidos pelo por R1 abaixo.
  - F3 Metadados incluem de maneira clara e explicita o identificador dos dados que eles descrevem
  - F4 Dados e metadados estão registrados ou indexados em fontes buscáveis

- *Accessible* (Acessível)

  Também é importante que todo dado localizável possua meios claros de acesso, que podem incluir autenticação e autorização

  - A1 Dados e metadados são recuperáveis pelos seus identificadores utilizando um protocolo de comunicações padronizado
    * A1.1 O protocolo é aberto, livre e universalmente implementável
    * A1.2 O protocolo permite processos de autenticação e autorização quando necessário
  - A2 Metadados são acessíveis, mesmo quando os próprios dados já não estão mais disponíveis

- *Interoperable* (Interoperáveis)

  Para a maior parte dos dados, é necessário que seja possível a integração com outros dados. Além disso, os dados precisam ter capacidade de serem interoperáveis com diferentes aplicações para análise, armazenamento e processamento

  - I1 Dados e metadados utilizam uma linguagem formal, acessível, compartilhada e amplamente aplicável para representação de dados
  - I2 Dados e metadados utilizam vocabulários que seguem, eles próprios, os princípios FAIR
  - I3 Dados e metadados incluem referencias qualificadas a outros dados e metadados

- *Reusable* (Reusáveis)

  O objetivo principal dessas regras é otimizar a reutilização de dados, sendo este portanto o principio final do conjunto.

- R1 Dados e metadados são ricamente descritos com uma pluralidade de atributos precisos e relevantes
  - R1.1 Dados e metadados são distribuídos com licenças de uso livres e acessíveis
  - R1.2 Dados e metadados são associados a uma procedência detalhada
  - R1.3 Dados e metadados atendem aos padrões da comunidade relevantes aos seus respectivos domínios

Há críticas relevantes a estes princípios, como por exemplo [BOECKHOUT; ZIELHUIS; BREDENOORD, 2018] que considera que eles não são suficientes para garantir formas responsáveis de compartilhamento de dados, que sua aplicação sem as qualificações devidas pode causar problemas adicionais, argumentando que é necessária sua suplementação com a boa governança das entidades produtoras destes dados. Contudo, um número crescente de organizações vem os adotando desde sua primeira publicação, sendo eles assim considerados uma das mais abrangentes formas de se garantir, e portanto adequadas para a plataforma que se deseja estabelecer.

2.2 ARM *(Atmospheric Research Measurement)*

O ARM *(Atmospheric Research Measurement)* é um programa iniciado pelo Departamento de Energia dos Estados Unidos em 1989 para monitorar a atmosfera terrestre. O objetivo do programa é prover para a comunidade científica observatórios estrategicamente posicionados e equipados com o estado-da-arte de sensoreamento para monitorar nuvens e aerossóis. Atualmente conta com três observatórios operacionais que buscam cobrir regiões de climas diversos: o observatório *Southern Great Plains* (SGP) está distribuído em varias localidades na região central do estado americano do Oklahoma e na região sul do estado do Kansas e busca estudar a atmosfera no meio-oeste dos Estados Unidos, o *Eastern North Atlantic* ENA está localizado no arquipélago de Azores e busca prover um dataset de longo prazo de nuvens em clima oceânico, consideradas uma das grandes formadoras de incertezas para modelos climáticos globais e regionais, por fim o *North Slope of Alaska* esta localizado no Alaska e busca estudar a atmosfera na região do Ártico. Embora o principal foco sejam os observatórios fixos, há também observatórios móveis capacitados para operar em qualquer ambiente, incluindo extremos de temperatura e umidade, com o intuito de capacitar campanhas de campo para coletar dados de regiões sub-representadas.

As informações encontradas no sistema de descoberta de dados ARM representam principalmente um ponto geográfico. Os pontos apresentados pelos dados são organizados em uma

hierarquia de dois níveis.

- 1 Locais: Um local é uma área geográfica onde a ARM coleta dados de forma independente ou em parceria com instituições locais.
- 2 Instalações: Uma instalação é caracterizada por grupos de instrumentos similares em diferentes pontos dentro do mesmo local.

Dada a quantidade significativa de informações recebidas pela ARM todos os dias, as fontes e usuários devem aderir aos padrões de dados da ARM. De acordo com os Padrões de Arquivos de Dados da ARM: Versão 1.2 [PALANISAMY, 2016], estabelecer padrões fornece consistência entre diferentes arquivos de dados, permitindo assim a reutilização de código com formatos distintos coerentes, arquivos de fácil leitura e programas simples para ler e tratar os arquivos padronizados.

Nota-se que conforme [PEPPLER et al., 2016] a preocupação com qualidade de dados está presente desde o início, com iniciativas como a checagem automática de consistência e desenvolvimento de experimentos de mensuração específica. Contudo, estes esforços foram irregulares e independentes, resultando em redundância e pouca capacidade de integração. Como resposta a isso, desde de 2000, o programa conta com um escritório de qualidade de dados, denominado *Data Quality Office* (DQO) localizado nas dependências da Universidade do Oklahama. Este escritório tem a responsabilidade de coordenar os esforços de qualidade de dados de todo o programa, e preparou os Padrões de Arquivo citado acima. Espera-se que a partir do desenvolvimento deste projeto, esforços similares possam ser estabelecidos para pesquisas atmosféricas no Brasil.

2.3 AmazonFACE

O AmazonFACE é um programa criado pela parceria entre pesquisadores de diversas instituições internacionais para estudar os efeitos das mudanças climáticas na Amazônia. Embora em alguns aspectos similar a outros do tipo, o diferencial deste programa é o estudo direto da influencia do enriquecimento de carbono em um ambiente controlado, ou *Free-Air CO2 Enrichment (FACE)* [RAMMIG; LAPOLA, 2022]. Para tanto, ele conta com a infraestrutura de torres instaladas em pontos estratégicos da floresta amazônica, se encontrando assim consideravelmente isolada e em locais de difícil acesso, como pode ser visto na figura 2.

Figura 2 – Torres usadas pelo AmazonFACE, ainda em fase de construção

Fonte: AmazonFACE.

Cada uma dessas torres contém um conjunto de sensores em certos níveis para monitorar os aspectos desejados em cada ponto. Na tabela 1 pode-se visualizar os sensores atualmente instalados no projeto e já capacitados para receber dados. Nota-se que essa lista foi feita pelo AmazonFACE, o que justifica seu uso da língua inglesa.

Type	**Model**	**Manufacturer (Geographical Info)**
Digital emispherical lens camera	Q25	Mobotix (Winnweiler, Germany)
Rain gauge	TB4	Hydrological Services Pty. Ltd. (Sydney, Australia)
Sunshine Pyranometer	SPN1	Delta-T Devices Ltd. (Burwell, Cambridge, UK)
Barometer	PTB101B	Vaisala Inc. (Vantaa, Helsinki, Finland)
Quantum	LI-190SB	LICOR Inc. (Lincoln, Nebraska, USA)
Ultrasonic Anemometer	WMT700	Vaisala Inc. (Vantaa, Helsinki, Finland)
Termohigrometer	HC2S3	Rotronic Instrument Corp. (Hauppauge, New York, USA)
Infrared Radiometer	SI-111	Apogee Instruments (Logan, Utah, USA)
Gas Analyzer	LI-840A	LICOR Inc. (Lincoln, Nebraska, USA)
Data Logger	CR1000	Campbell Scientific (Logan, Utah, USA)
Infrared gas analyser	LI-6800 F	LICOR Inc. (Lincoln, Nebraska, USA)
Minirhizotron camera	BTC2	Bartz Technology (Ventura, California, USA)
Soil CO2 Flux System	LI-8100A	LICOR Inc. (Lincoln, Nebraska, USA)
Stand-Alone Logging Dendrometer	DBL60	ICT International (Armidale, NSW, Australia)
Sap Flow Meter	SFM1	ICT International (Armidale, NSW, Australia)
Profile Probe	PR2/6	Delta-T Devices Ltd. (Burwell, Cambridge, UK)
Infrared gas analyser	EGM-4	Environmental & Gas Monitoring Ltd (Galston, UK)
Leaf Porometer Stomatal Conductance	SC-1	Decagon Devices, Inc. (Pullman, Washington, USA)

Tabela 1 – Lista de Instrumentos, conforme fornecida pelo AmazonFACE

A partir disso, ele busca responder a seguinte pergunta: ”Como o aumento da concentração de CO2 na atmosfera afeta a biologia e a bioquímica da floresta Amazônica, a biodiversidade contida nela e o serviços de eco-sistema que ela provém”.

Este projeto atualmente possui problemas significativos de governança de dados, enfrentando perdas frequentes e não possuindo ciclos de vida bem estabelecidos para todos os dados coletados. Além disso, os efeitos da pandemia de COVID-19 dificultaram ainda mais o trabalho de criação de uma estrutura para solucionar estes problemas, produzindo atrasos e períodos longos sem que fosse possível realizar o monitoramento adequado.

Os dados atualmente disponibilizados pelo AmazonFACE estão divididos seguindo a lógica padrão de [GIEBLER et al., 2020], divididas portanto em zonas.

As zonas desempenham um papel crucial na organização e no processamento dos dados, fornecendo uma estrutura que permite o gerenciamento eficiente da informação. Nesse contexto, a divisão tradicional em zonas *“raw”*, *“staging”* e *“trusted”* representa uma abordagem estratégica para a manipulação de dados em diferentes estágios do ciclo de vida.

A zona *“raw”* é o ponto de entrada inicial para os dados. Nessa fase, os dados são armazenados exatamente como são recebidos, sem transformações significativas. Os arquivos estão, portanto, em formato binário e ASCII. Isso preserva a integridade original dos dados brutos, possibilitando análises futuras sem perder detalhes essenciais. A zona *“raw”* serve como um repositório central e flexível para dados de diversas fontes.

À medida que os dados são movidos para a próxima fase, a zona *“staging”*, ocorre um processo de organização e limpeza preliminar. Nessa etapa, os dados são estruturados e formatados de maneira consistente, em formato CSV, eliminando inconsistências e redundâncias. A zona *“staging”* funciona como uma área intermediária onde os dados são preparados para serem consumidos de maneira mais eficiente nas etapas subsequentes.

Finalmente, a zona *“trusted”* representa a camada mais refinada e confiável. Nessa fase, os dados passam por transformações avançadas, como enriquecimento, agregação e aplicação de análises sistemáticas. A ênfase é na criação de uma fonte de dados confiável e pronta para análises detalhadas e decisões críticas. A zona *“trusted”* garante a integridade, a qualidade e a segurança dos dados, tornando-se a base para análises avançadas e relatórios críticos. Nenhum dado do projeto está atualmente nesta zona.

Pela facilidade de submissão e adequação com os frameworks utilizados, optou-se por priorizar a implementação das capacidades da plataforma sobre o conjunto de dados da zona *“staging”*.

2.4 Formatos de dados

Um dos aspectos a serem levados em consideração para este projeto é o formato dos dados que serão trabalhados pelas ferramentas desenvolvidas. Algumas opções possíveis são seguir o padrão mais próximo possível do ARM, baseado em NetCDF, ou buscar outros padrões comumente usados em pesquisa, como CSV, JSON e XML.

O NetCDF é o formato padrão utilizado pelos observatórios ARM, tendo sido desenvolvido para suportar criação, acesso e compartilhamento de dados científicos ordenados. O formato é mantido pela *University Corporation for Atmospheric Research* (UCAR) e possui como vantagens a capacidade de auto-descrição, a partir de um cabeçalho e independência da plataforma de uso, possuindo bibliotecas específicas para lidar com a portabilidade. Além disso, há o compromisso pelo suporte a todas as versões presentes e futuras do formato, fornecendo assim garantias de arquivamento fundamentais para os objetivos buscados. Dessa forma, ele permite que metadados e dados sejam distribuídos no mesmo arquivo. Os arquivos manipulados pela ARM são nomeados de acordo com a convenção descrita pelo Comitê de Padrões da ARM [PALANISAMY, 2016], que são organizados da seguinte forma:

(sss)(inst)(qualifier)(temporal)(Fn).(dl).(yyyymmdd).(hhmmss).nc

- (sss): 3 letras que identificam o local dos dados
- (inst): abreviação do nome do instrumento
- (qualifier): qualificador opcional que distingue o arquivo dos demais produzidos pelo mesmo instrumento no mesmo local
- (temporal): descrição opcional da resolução temporal do arquivo, seguida pela unidade utilizada (por exemplo, 30 m, 1 h, 200 ms, 14 d)
- (Fn): designação da instalação da ARM
- (dl): indicador de nível de dados, 00 ou uma letra minúscula seguida pelo número do nível de dados
- (yyyymmdd): coordenada UTC que indica a data da primeira medição do arquivo
- (hhmmss): horário UTC, horas, minutos e segundos
- .nc: extensão de arquivo para o formato netCDF

Esse padrão facilita o entendimento do dado antes mesmo que o arquivo seja aberto, mas também exige um alto nível de conhecimento prévio do pesquisador que irá manipula-lo, dificultando sua implementação prática.

Por outro lado, uma opção mais favorável para um desenvolvimento inicial e amigável a pesquisadores pouco experientes é o CSV.

O formato CSV *(Comma-Separated Values)* é uma representação textual de dados tabulares amplamente utilizado na ciência de dados. Nesse formato, as informações são organizadas em linhas, sendo cada linha uma entrada separada por vírgulas. Cada entrada pode representar um registro de dados, em diversos formatos possíveis (datas, inteiros, pontos flutuantes, etc). O CSV é valorizado pela sua simplicidade e interoperabilidade, sendo facilmente legível por humanos e facilmente processado por máquinas. Essa versatilidade torna-o uma escolha comum para troca de dados entre diferentes sistemas e ferramentas de análise, facilitando a importação e exportação de conjuntos de dados em uma variedade de contextos . Também há a vantagem de o AmazonFACE já utilizar esse formato em seus processos, contribuindo para integração e testes preliminares da plataforma. Outros formatos possíveis, como JSON e XML, também são suficientemente simples para serem implementados em uma plataforma inicial, mas não possuem esta vantagem.

Espera-se que a escolha pelo CSV permita, dessa forma, seguir um padrão consistente para o uso de arquivos neste projeto.

2.5 Fluxos de dados

Resultados bem sucedidos em pesquisa dependem fundamentalmente de fluxos de dados eficientes para realizar análises robustas e extrair insights significativos. Esses fluxos envolvem a coleta, transformação, armazenamento e análise de dados ao longo de todo o ciclo de vida do projeto.

A primeira etapa, coleta de dados, abrange a aquisição de informações de diversas fontes, como bancos de dados e sensores, entre outros. Em seguida, ocorre a fase de transformação, na qual os dados são limpos e preparados para análise. A armazenagem adequada desses dados é crucial e geralmente realizada em bancos de dados, data lakes ou outras infraestruturas de armazenamento escaláveis.

A etapa final, análise de dados, é onde cientistas e analistas exploram os dados para identificar padrões, tendências e informações relevantes. Além disso, também é interessante nesta fase oferecer ferramentas para visualização e interpretação dos resultados.

É nessa etapa que se planeja implementar as principais funcionalidades desenvolvidas nesse trabalho, permitindo que, a partir de boas informações de qualidade, seja impulsionada a geração de valor científico.

3 PROJETO DA ARQUITETURA

3.1 Metodologia de Trabalho

A metodologia utilizada no desenvolvimento deste trabalho foi elaborada com base nas necessidades de pesquisadores do AmazonFACE e de discussões com o orientador e outros colaboradores. A partir disso, pode-se definir as tecnologias necessárias para permitir não apenas o sucesso do projeto atual, como também de futuros desenvolvimentos a partir deste.

A metodologia seguirá um passo a passo que permitirá sua implementação gradual, com testes em cada etapa, para garantir o funcionamento correto e a integração entre as partes.

- Streamlit

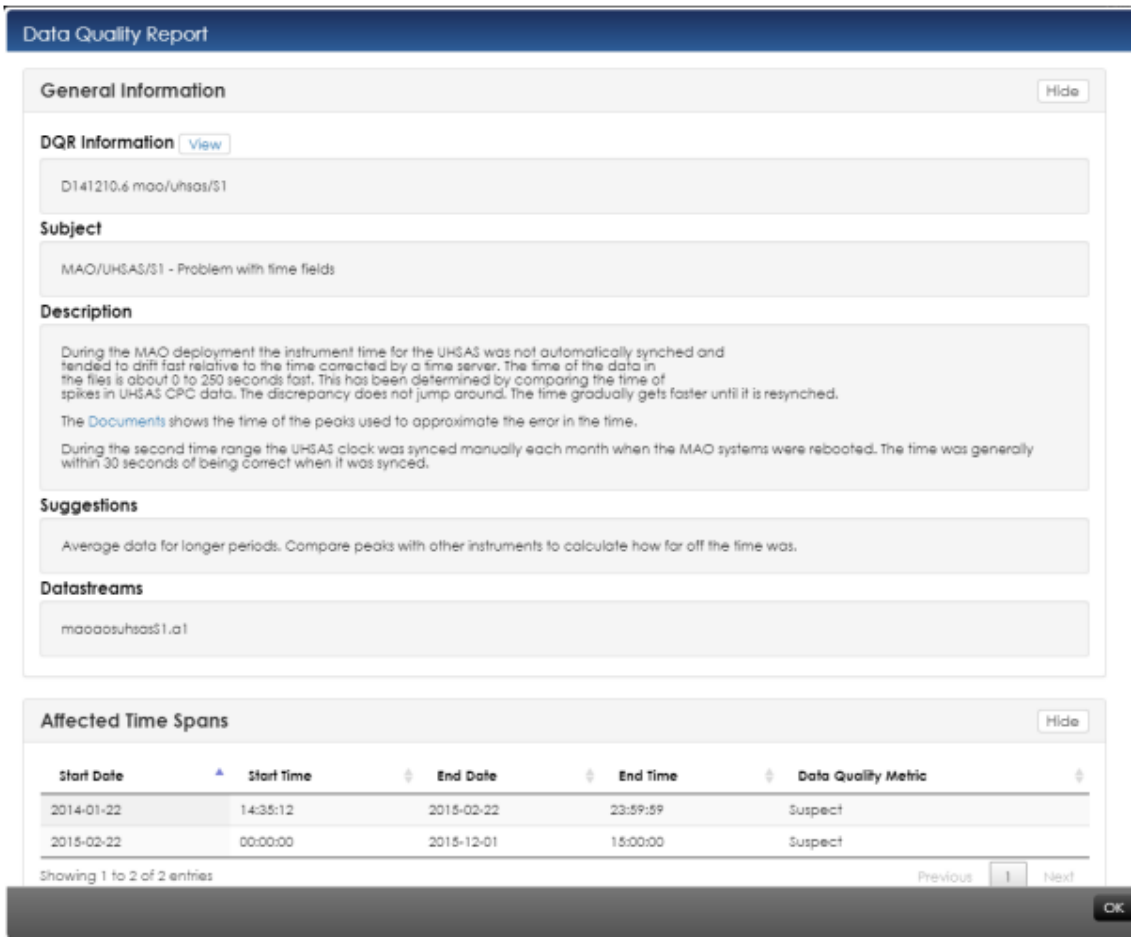
  O framework escolhido para o desenvolvimento da interface. Alguns recursos e capacidades que influenciaram nesta decisão foram: a facilidade de uso, permitindo a criação de novos elementos com poucas linhas de código; capacidade de rápida prototipagem e modificação; e existência de componentes integrados para visualização e exploração de dados.

  Juntando-se todos estes fatores, foi considerado o framework ideal para este projeto

- Data Quality Report (DQR)

  O Data Quality Report é uma das funcionalidades mais importantes da interface, sendo nela que o pesquisador irá informar sobre a qualidade dos dados inseridos na plataforma. Conforme descrito anteriormente, muito de sua implementação será baseada nas capacidades já estabelecidas no ARM, com adaptações acrescentadas conforme as necessidades identificadas. A figura 3 ilustra um exemplo de Data Quality Report extraído da plataforma ARM.

Figura 3 – Exemplo de Data Quality Report utilizado na plataforma ARM



Fonte: ARM.

Algumas informações relevantes que podem ser vistas são detalhes do instrumento de onde partiram os dados, período de coleta, uma breve descrição dos problemas encontrados (caso existam) e uma sugestão, caso possível, para corrigir esses problemas e permitir a utilização dos dados.

- Bases de dados não relacionais e MongoDB

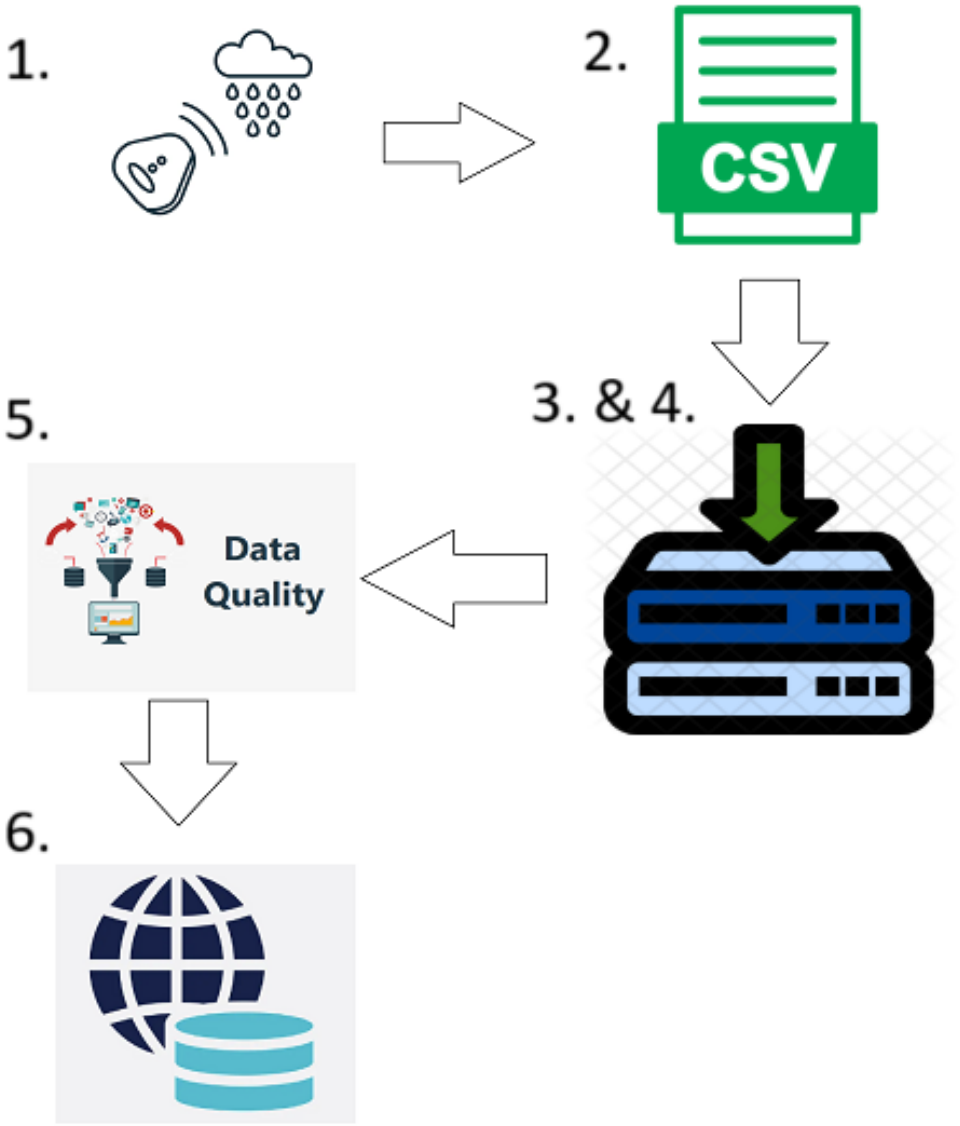
  A base de dados é um foco importante do projeto, especialmente considerando-se os já citados problemas frequentes com perdas. A escolha de um bom framework para o armazenamento leva em consideração a flexibilidade, escalabilidade e compatibilidade da plataforma. Bases de dados baseadas em MongoDB são adequadas a esse propósito, sendo este um framework orientado a documentos e possuindo as capacidades de modificação e expansão requeridas do sistema, sem o uso forçado de modelos de dados específicos.

- *Fluxo de dados*

  Utilizando-se de todas as ferramentas desenvolvidas, pode-se estabelecer o fluxo de dados

esperado para um pesquisador envolvido no projeto, como visto na figura 4.

Figura 4 – Fluxo de Dados esperado



Fonte: Elaboração própria.

1. Coleta de dados pelo gestor do equipamento
2. Transformação dos dados em formato CSV para ser recebido pela plataforma
3. Submissão dos dados utilizando interface desenvolvida, no site do projeto
4. Dados em *"standby"* até o próximo passo
5. Preenchimento do Data Quality Report pelo pesquisador
6. Disponibilização dos dados no portal (incluindo Data Quality Report)

3.2 Especificação de Requisitos

Neste capítulo são apresentados os requisitos funcionais que determinam as capacidades esperadas do projeto bem como os requisitos não funcionais que estabelecem a metodologia e tecnologia a ser utilizada na implementação do projeto.

3.2.1 Requisitos Funcionais

Após levadas em conta todas as definições, padrões e tecnologias a serem utilizadas no projeto, pode-se definir os requisitos funcionais que se espera que esta plataforma seja capaz de suprir.

Capacidade de criação de novas Campanhas: O usuário cadastrado pode criar uma campanha com período definido para organizar seus dados

Capacidade de inserção e recuperação de dados científicos em formato padronizado: é esperado que todos os pesquisadores possam inserir seus dados após o processamento devido.

Capacidade de registro da qualidade dos dados inseridos: o Data Quality Report desenvolvido deve permitir que todos os dados, independentemente de sua origem, possa receber uma análise de qualidade descritiva e completa.

Capacidade de acesso aos dados de outros pesquisadores: evidentemente, todos os dados registrados devem estar acessíveis dentro da plataforma e disponibilizados para download.

Nota-se que a maioria destas possui paralelos com as capacidades encontradas no portal do ARM, adaptadas para o contexto delimitado.

Outro aspecto relevante é que embora esta necessidades sejam consideradas as mínimas suficientes para um produto final aplicável, é também levada em conta a capacidade necessária para eventual expansão do projeto, de forma que futuras exigências possam ser facilmente implementadas.

3.2.2 Requisitos Não Funcionais

Os requisitos não funcionais definem expectativas para a entrega do sistema, devendo ser levados em conta durante todos os passos de desenvolvimento para garantir que as funcionalidades desenvolvidas não sejam desperdiçadas. Abaixo estão elencados os requisitos

Desempenho: Para as capacidades iniciais do sistema, considera-se necessário que seja

consideravelmente leve e capaz de ser implementado em máquinas comerciais simples. Eventualmente, espera-se que seja possível sua implementação em nuvem.

Escalabilidade: Apesar do funcionamento inicial focado em desempenho, espera-se que o sistema tenha capacidade de ser escalável, visto que o registro futuro de grandes volumes de dados de diferentes formatos é planejado para fases futuras do projeto.

Segurança: As necessidades de segurança do projeto serão supridas pelo autenticação e autorização de usuários integradas em sistemas próprios, garantindo ao administrador controle sobre quem irá manipular o sistema.

Facilidade de uso: Dados os objetivos amplos da plataforma, é esperado que qualquer pesquisador possa facilmente se familiarizar com as funcionalidades do sistema.

3.3 Desenvolvimento da Arquitetura

Nesta seção, será explicitada a arquitetura desenvolvida a partir dos conceitos discutidos, incluindo decisões de projeto relevantes e diagramas explicativos. A implementação a partir desse levantamento será discutida no capitulo 4.

3.3.1 Elaboração da Arquitetura de Interações

As funcionalidades e interações que o usuário poderá realizar foram definidas a seguir. Nota-se que o apresentado nesta etapa se refere ao considerado o mínimo necessário para a Prova de Conceito desenvolvida, sendo portanto considerado um trabalho futuro a implementação de novas funcionalidades.

Conforme as necessidades observadas, tem-se 3 fluxos de interações principais na plataforma.

Criação de Nova Campanha: Após o usuário completar seu registro ou login, ficará disponível uma tela em que ele poderá criar novas campanhas, além de interagir e revisar campanhas já submetidas. Para criar uma nova campanha, será necessário preencher um formulário contendo as informações relevantes, que será explorado mais adiante. Por fim, ele pode optar ou não por imediatamente inserir dados nesta campanha.

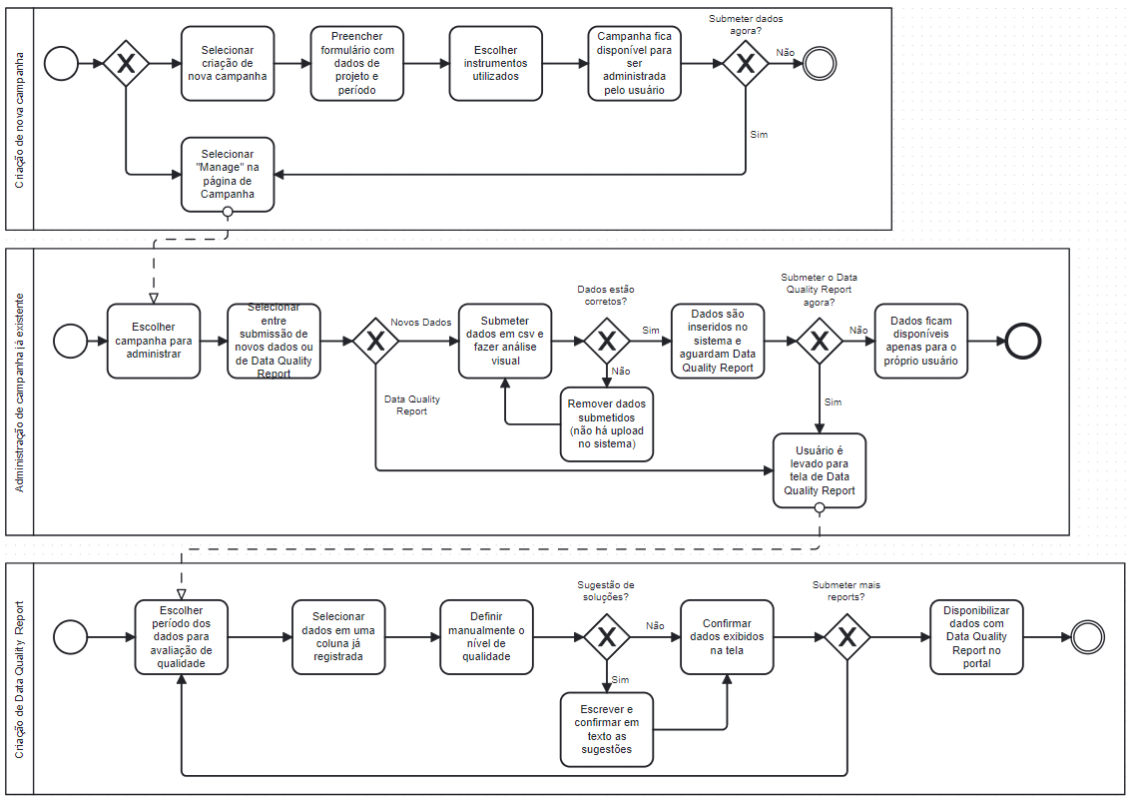
Administração de campanha já existente: Após criar uma campanha, um usuário pode submeter novos dados a ela, ou criar um Data Quality Report caso já tenha submetido dados anteriores. Os dados passam por uma etapa de verificação antes da submissão final e, caso esta seja

realizada, são armazenados e então aguardam a submissão de informações de qualidade.

Criação de Data Quality Report: Por último, após a campanha já possuir dados inseridos, é possível criar um ou mais Data Quality Reports, que irão descrever a qualidade dos dados e disponibiliza-los para demais usuários. Vale ressaltar que esta criação pode ser feita a qualquer momento após a inserção dos dados, mesmo decorrendo períodos longos de espera, mas os dados só ficarão disponíveis para outros usuários depois de sua conclusão.

Todos os fluxos são encerrados em alguma submissão de novos dados - dados de Campanha, de pesquisa e de qualidade, respectivamente -, e são sequenciais: em outras palavras, só é possível Administrar uma campanha após sua criação, e só é possível criar um Data Quality Report após a submissão de dados. Na figura 5 pode ser visto um diagrama BPMN *(Business Process Model and Notation)* indicando este processo por inteiro.

Figura 5 – BPMN da Plataforma completa

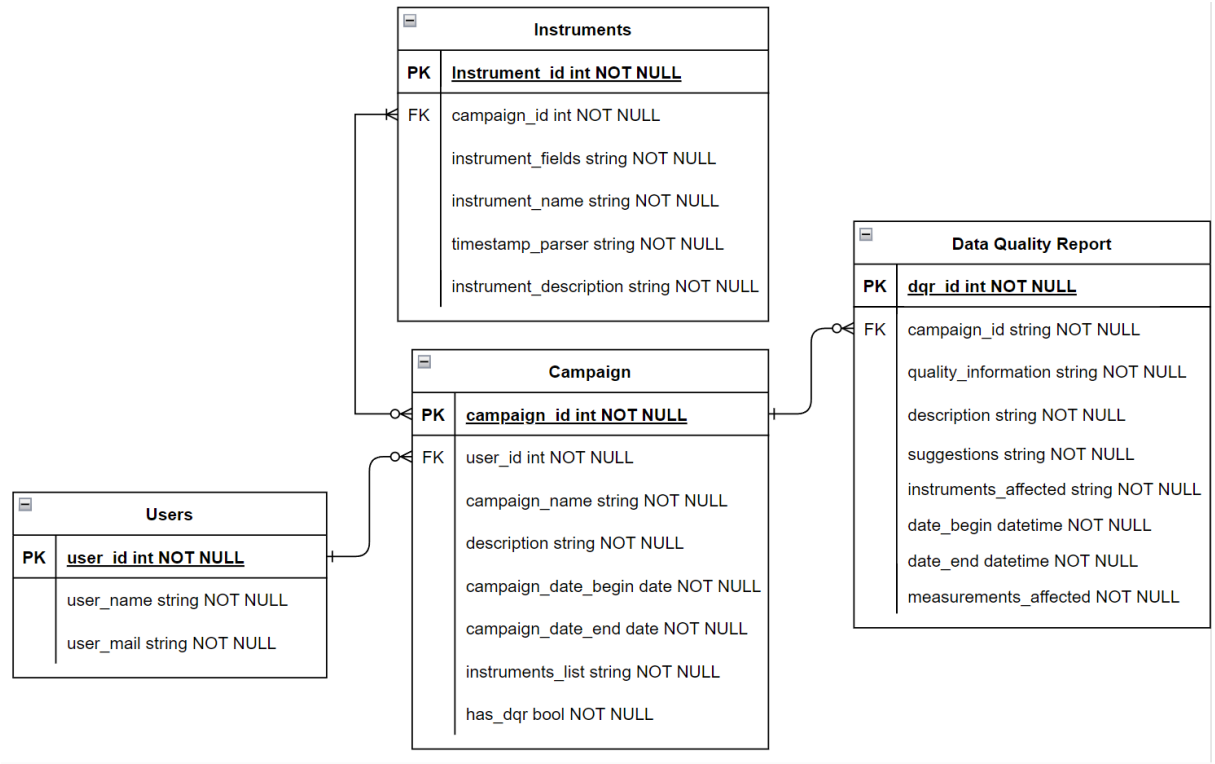


Fonte: Elaboração própria

3.3.2 Elaboração do Modelo de Dados

A partir das necessidades da plataforma, foi verificado a necessidade da criação de múltiplas entidades, que podem ser vistas por completo na figura 6:

Figura 6 – Entidades do Sistema



Fonte: Elaboração própria.

Explorando em mais detalhes cada uma das entidades, pode-se começar pelo User. Essa é a principal entidade de controle do sistema, que irá de fato interagir com as telas descritas na seção 3.3.1, criando novas campanhas ou administrando campanhas já existentes, submetendo dados e criando Data Quality Reports. A tabela 2 detalha todos os campos desta entidade.

Tabela 2 – Elemento User

Campo	**Tipo**	**Descrição**
user_id	int	id do usuário
username	string	nome do usuário
usermail	string	email do usuário

Fonte: Elaboração própria.

Campanha é a classe que registra as informações temporais e de instrumentação, indicando a fonte dos dados criados e organizando-os dentro da plataforma. Ela também identifica se já foi ou não submetido um Data Quality Report dos dados inseridos. A tabela 3 detalha todos os campos desta entidade.

Tabela 3 – Elemento Campaign

Campo	**Tipo**	**Descrição**
campaign_id (chave privada)	int	id da campanha
user_id(chave estrangeira)	int	id do usuário
campaign_name	string	nome da campanha
description	string	descrição da campanha
campaign_date_begin	date	data de início da campanha
campaign_date_end	date	data de fim da campanha
instruments_list	string	lista de instrumentos na campanha
has_dqr	bool	existencia de data quality report na campanha

Fonte: Elaboração própria.

A entidade de instrumentos contém os dados de cada instrumento registrado no sistema, incluindo nome, descrição, campos coletados e formato em que os registros temporais são salvos. A tabela 4 detalha todos os campos desta entidade.

Tabela 4 – Elemento Instruments

Campo	**Tipo**	**Descrição**
instrument_id (chave privada)	int	id do instrumento
campaign_id (chave estrangeira	int	id da campanha
instruments_fields	string	campos coletados no instrumento
instruments_name	string	nome do instrumento
timestamp_parser	bool	formato do registro temporal do instrumento
instruments_description	string	descrição do instrumento

Fonte: Elaboração própria.

Por fim a entidade Data Quality Report contém as informações relevantes de qualidade dos dados submetidos, como nível de qualidade e descrição dos problemas. Uma visão mais abrangente será vista mais adiante, no capítulo 4.Implementação. Vale ressaltar que, embora os dados inseridos não possuam formato fixo, os Data Quality Reports em si devem ser padronizados, para permitir um controle metódico da qualidade. A tabela 5 detalha todos os campos desta entidade.

Tabela 5 – Elemento Data Quality Report

Campo	**Tipo**	**Descrição**
dqr_id (chave privada)	int	id da Data Quality Report
campaign_id(chave estrangeira)	int	id da campanha
quality_information	string	informação do nivel de qualidade
description	string	descrição do problema de qualidade
instruments_list	string	lista de instrumentos afetados pelo problema
date_begin	date	data de início do problema
date_end	date	data de fim do problema
measurements_affected	string	lista de medições afetadas pelo problema

Fonte: Elaboração própria.

Sobre as relações entre as entidades temos que um usuário pode criar múltiplas campanhas, mas todas estarão exclusivamente associadas a ele. Da mesma forma, uma campanha poderá conter múltiplos instrumentos (como pluviômetros, termômetros, barômetros, etc.), mas cada instrumento poderá estar presente em múltiplas campanhas, ou em nenhuma. Por fim, cada Data Quality Report estará exclusivamente associado a uma campanha, que por sua vez poderá ter múltiplos deles associados a ela.

Nota-se dois grupos relevantes que não possuem entidade: usuários não registrados, visto que não é exigida deles nenhuma informação para que tenham acesso aos dados disponíveis na plataforma; e os próprios dados inseridos, visto que o formato deles não é padronizado e portanto não pode ser adequadamente modelado dentro do DER.

3.3.3 Elaboração da Infraestrutura Computacional

A arquitetura de infraestrutura é monolítica, caracterizada por sua estrutura integrada e centralizada, que continua a ser uma abordagem sólida no desenvolvimento de software. [BLINOWSKI; OJDOWSKA; PRZYBYŁEK, 2022] A principal vantagem da arquitetura monolítica reside em sua simplicidade. Ao consolidar todos os componentes em um único código-fonte, simplifica-se o processo de desenvolvimento. Essa característica facilita a integração de novos desenvolvedores e contribui para uma curva de aprendizado adequada. Além disso, em termos de manutenção contínua, a coesão entre os módulos em uma arquitetura monolítica simplifica consideravelmente este processo. Atualizações e correções podem ser implementadas de maneira direta, sem a necessidade de coordenação complexa entre diferentes partes do sistema. Isso contribui para uma gestão mais eficiente das mudanças no software. A concentração de

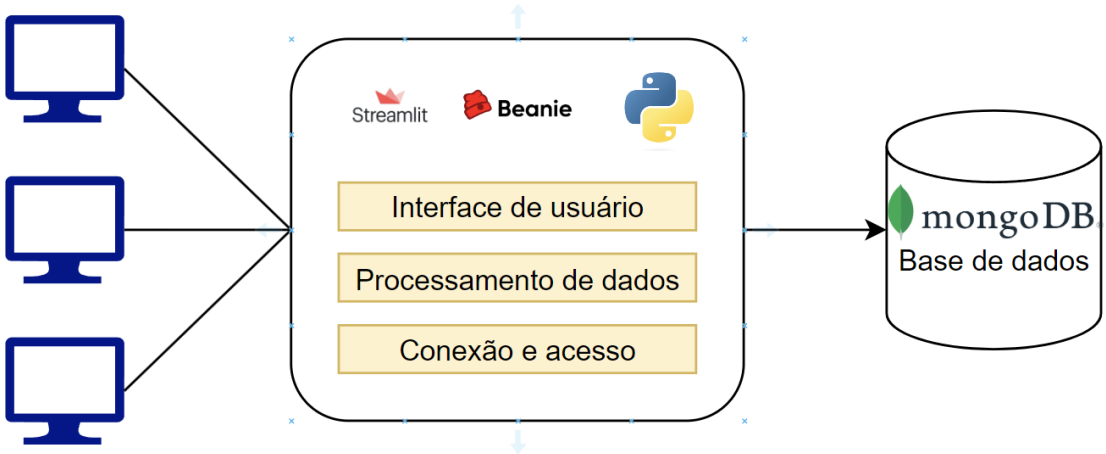
todo o código em um ambiente único também facilita o processo de identificação e correção de bugs. O debugging se torna mais direto, reduzindo a complexidade associada ao rastreamento de erros em sistemas distribuídos.

Por fim, também há vantagens na eficiência: a execução de uma aplicação monolítica muitas vezes resulta em melhor desempenho. A comunicação direta entre os componentes contribui para uma eficiência em cenários nos quais a escalabilidade horizontal não é uma prioridade imediata. A implementação de uma arquitetura monolítica muitas vezes implica custos iniciais inferiores. Requer menos infraestrutura e configuração, resultando em uma redução nos custos iniciais. Além disso, a complexidade geral do sistema é diminuída, o que facilita a implementação e a manutenção, especialmente em estágios iniciais do desenvolvimento.

Embora não seja uma solução universal para projetos de software, a arquitetura monolítica oferece uma abordagem eficaz para este cenário, proporcionando simplicidade, eficiência e facilidade de manutenção.

Dessa forma, ela foi escolhida para este projeto, implementando aspectos como a interface do usuário, com as interações da seção 3.3.1 e as telas do próximo capítulo; processamento de dados, com a organização de arquivos e geração de tabelas que será vista mais adiante; e conexão com o banco de dados; de maneira integrada em um único ponto. Assim, frameworks como o Streamlit, Beanie (para comunicação com MongoDB), e tecnologias adjacentes ao Python são combinadas para gerar o mostrado na figura 7.

Figura 7 – Arquitetura monolítica da plataforma



Fonte: Elaboração própria

4 IMPLEMENTAÇÃO

Este capítulo apresenta o produto implementado, após todo o projeto apresentado e descrito no capítulo anterior que fora projetado e descrito ao longo do capítulo anterior. Será mostrado como foi implementada a plataforma, incluindo todas as telas relevantes a serem preenchidas por um usuário. O código desenvolvido pode ser encontrado em https://github.com/GandraOProprio/DataQualityDataMap [GONCALVES, 2023].

.

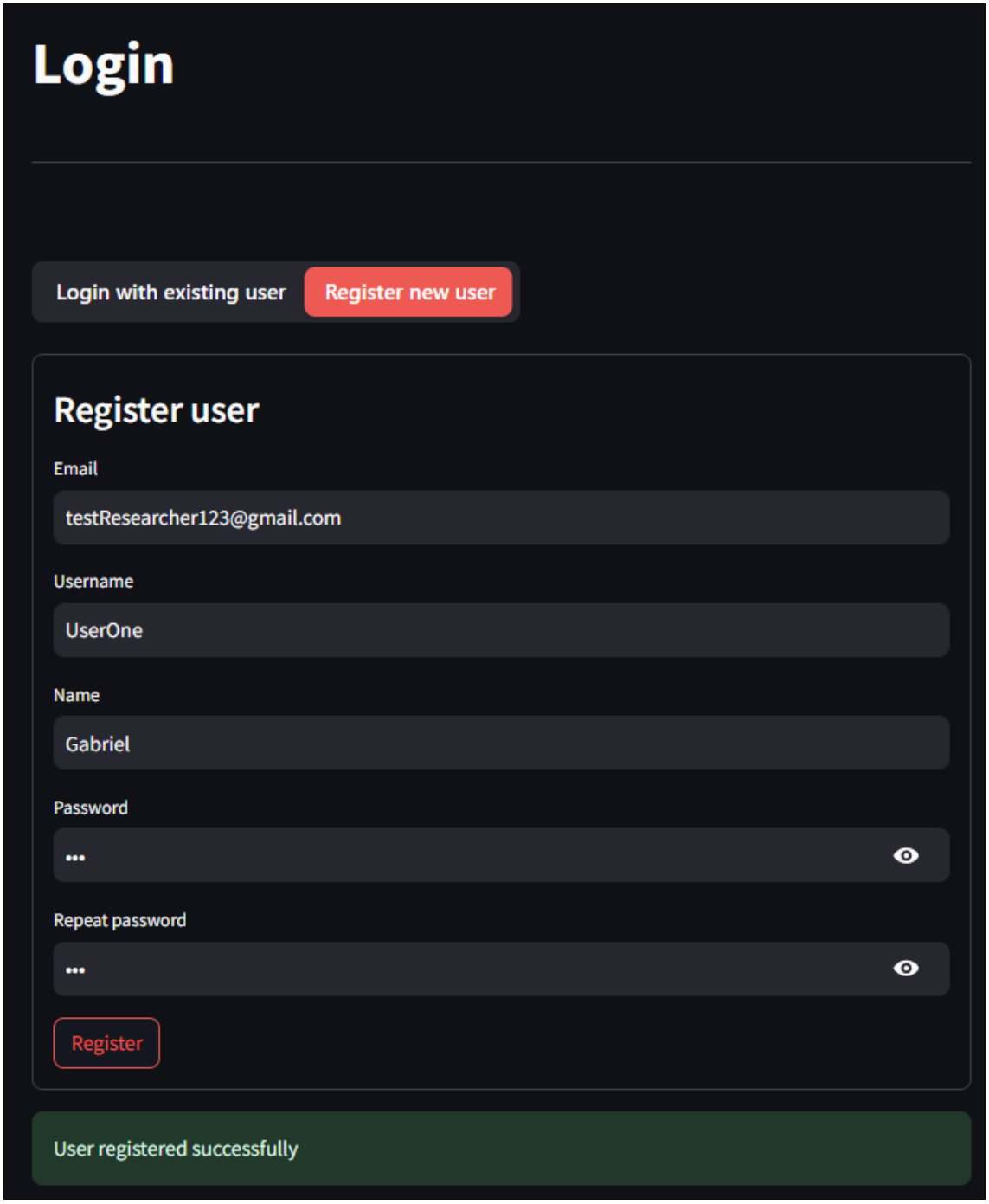
Nesse ponto, uma decisão de projeto importante de ser levantada foi a opção por desenvolver a plataforma na língua inglesa. Isso ocorre tanto pelas próprias características do Amazon-FACE, que é um projeto internacional com muitos de seus dados e planejamentos desenvolvidos nesta língua, quanto pela ambição de expandir e transformar as ferramentas apresentadas aqui em um projeto maior, que inclua pesquisadores de muitas nacionalidades.

As seções seguintes apresentam todo o projeto e funcionam apoiadas sobre a arquitetura mostrada.

4.1 Registro de Usuários

O cadastro de novos usuários se dá utilizando uma biblioteca própria do streamlit, denominada streamlit_authenticator. Essa biblioteca inclui diversas funcionalidades relevantes, dentre as quais as que foram utilizadas foram a capacidade de registrar novos usuários e a capacidade de realizar login por usuários cadastrados. Por outro lado, funcionalidades de recuperação de senha e mudança de nome não foram consideradas essenciais, e não chegaram a ser implementadas. Para os fins da plataforma desenvolvida, foi permitido que qualquer pessoa pudesse realizar o cadastro, mas o administrador do sistema pode optar por limitar essa permissão para pesquisadores de uma lista pré-autorizada. Um exemplo de registro de novo usuário pode ser visto na figura 8.

Figura 8 – Registro de usuário bem sucedido



Fonte: Elaboração própria.

Um fator relevante dessa biblioteca é que há um algoritmo de hash automático no momento em que o usuário é registrado, de forma que nunca há senhas salvas em *plain text* e cumprindo com os princípios da segurança cibernética. Outro ponto importante é uma função adicional acrescentada de manter o usuário registrado, mesmo quando a página é recarregada. Esta capacidade vai além do estabelecido no session_state, e se baseia no uso de cookies com um tempo de expiração também determinado pelo administrador (e por padrão, definido em 30 dias), evi-

tando assim necessidade de conexões sucessivas dentro desse período. A tela de usuário, que eventualmente poderia conter

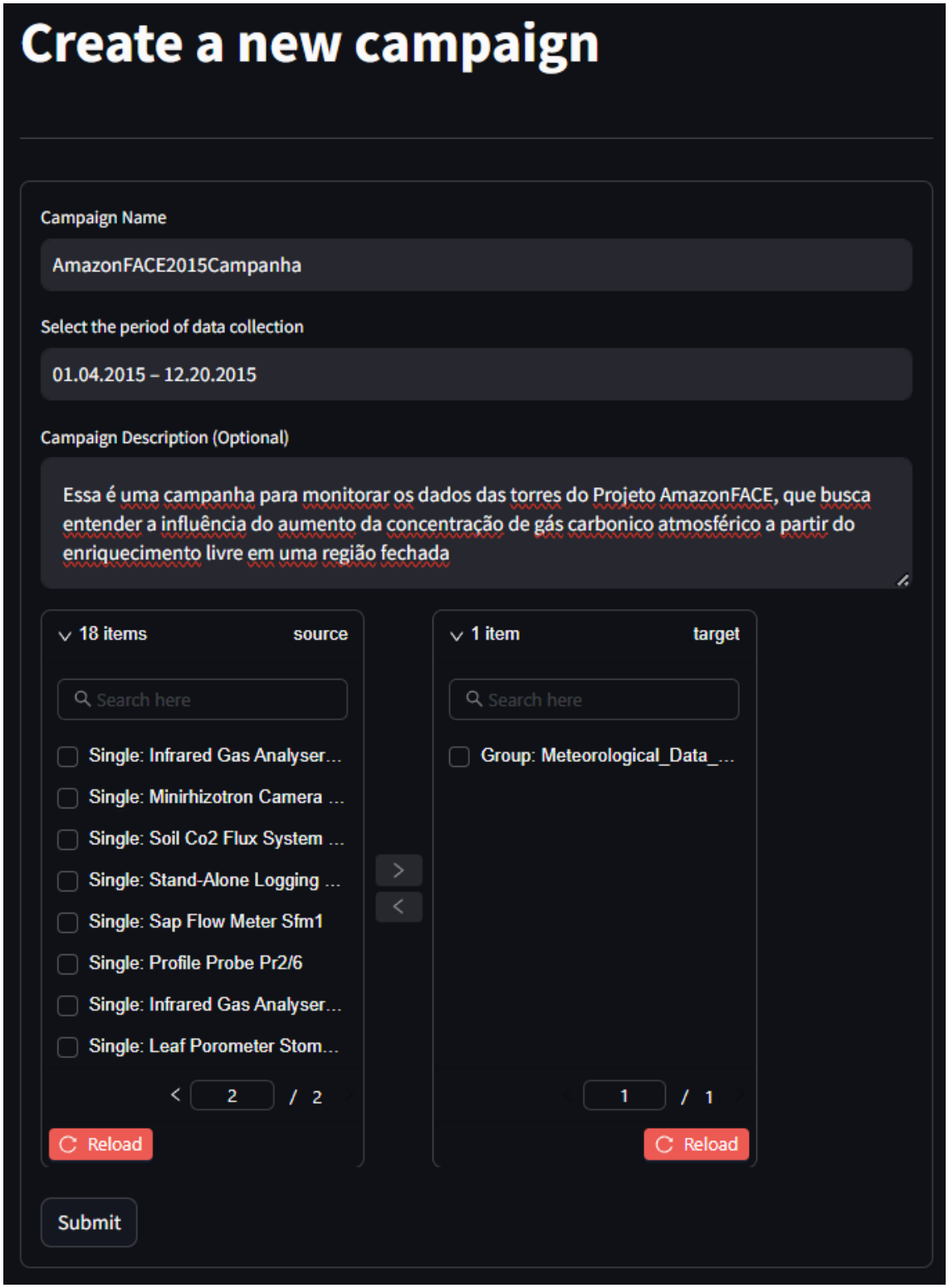
É importante ressaltar novamente que, exceto pela página inicial, que contém a opção de download de todos os dados que possuem Data Quality Report conforme os princípios estabelecidos em seções anteriores, nenhuma outra página da plataforma pode ser acessada sem a autenticação do usuário, permitindo assim um elevado grau de controle sobre quem está inserindo dados na plataforma. Como medida adicional, o streamlit também detecta e impede o processamento de arquivos fora do formato .csv, evitando danos causados por tentativas de submeter arquivos maliciosos.

4.2 Criação de Campanha

A campanha atua como a principal divisão para os dados que serão inseridos na plataforma, servindo para diferenciá-los entre si mesmo que possuam perfil similar, como é o caso de dados coletados por um mesmo conjunto de instrumentos em anos diferentes.

Os principais fatores que a definem são possuir início e fim bem definidos, e um conjunto de instrumentos associados, a partir de uma base pré-estabelecida. Para diferenciá-la de outras campanhas, é sugerido um nome único, preferencialmente relacionado com o alvo de estudo. Também é recomendado que o usuário sempre forneça uma breve descrição, indicando por exemplo objetivos científicos buscados e resultados esperados, embora esta não seja uma parte obrigatória do registro. Na figura 9 podemos ver um exemplo prático desse cadastro.

Figura 9 – Cadastro de campanha bem sucedido



Fonte: Elaboração própria.

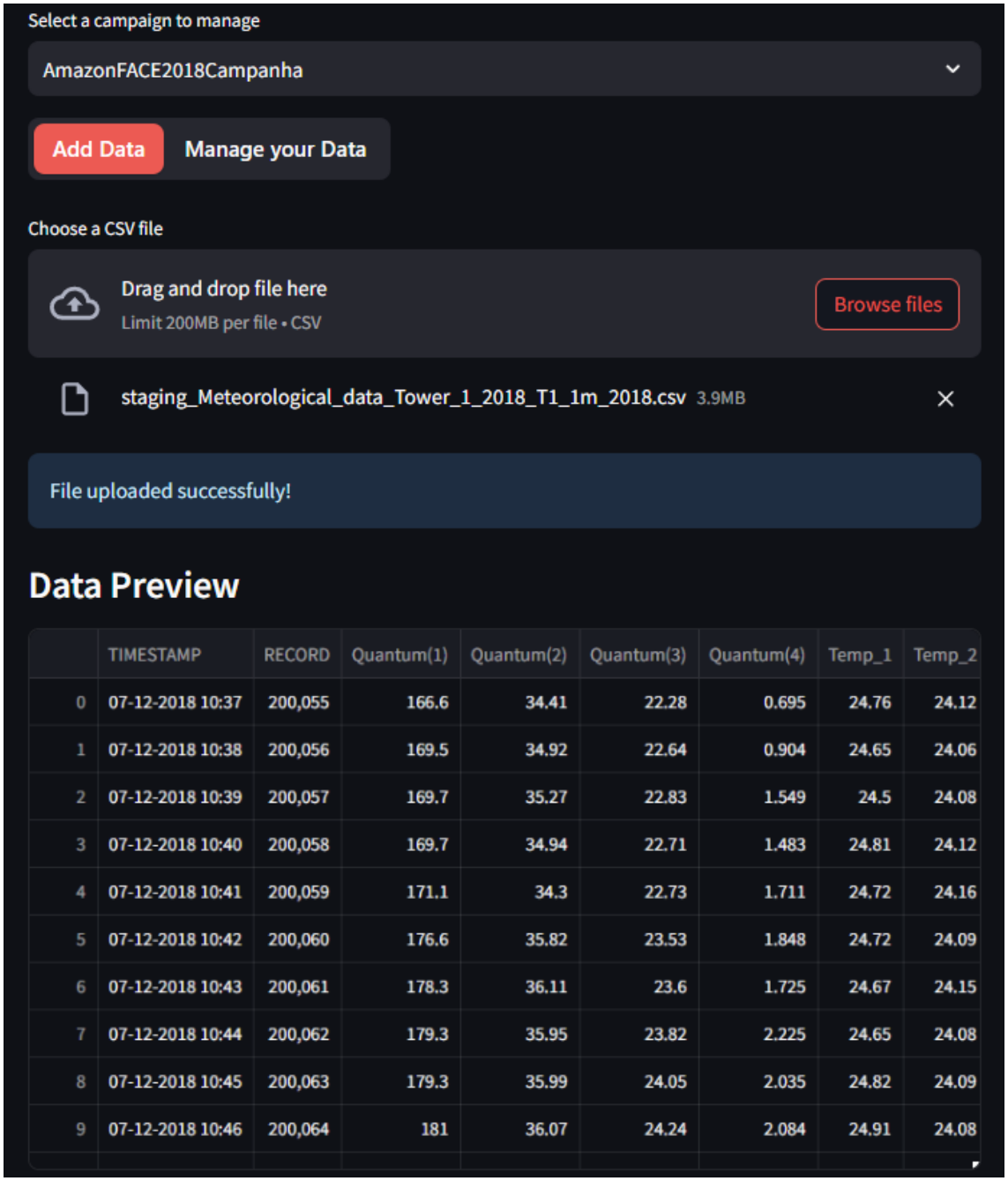
Em termos de consistência, embora o sistema seja projetado para receber um conjunto altamente variado de dados, é esperado que dentro de uma mesma campanha possuam um mesmo

formato, sendo essa portanto a única limitação relevante levantada nessa etapa. Ao criar uma campanha, é gerado também um identificador único para ela, garantindo que usuários distintos que eventualmente venham a criar campanhas com um mesmo nome não tenham problemas com conflitos.

4.3 Submissão de dados na plataforma

A submissão de dados na plataforma, após a criação de uma campanha, foi projetada para ser feita de forma intuitiva. Na figura 10 podemos ver um exemplo da inserção inicial realizada.

Figura 10 – Inserção de dados



	TIMESTAMP	RECORD	Quantum(1)	Quantum(2)	Quantum(3)	Quantum(4)	Temp_1	Temp_2
0	07-12-2018 10:37	200,055	166.6	34.41	22.28	0.695	24.76	24.12
1	07-12-2018 10:38	200,056	169.5	34.92	22.64	0.904	24.65	24.06
2	07-12-2018 10:39	200,057	169.7	35.27	22.83	1.549	24.5	24.08
3	07-12-2018 10:40	200,058	169.7	34.94	22.71	1.483	24.81	24.12
4	07-12-2018 10:41	200,059	171.1	34.3	22.73	1.711	24.72	24.16
5	07-12-2018 10:42	200,060	176.6	35.82	23.53	1.848	24.72	24.09
6	07-12-2018 10:43	200,061	178.3	36.11	23.6	1.725	24.67	24.15
7	07-12-2018 10:44	200,062	179.3	35.95	23.82	2.225	24.65	24.08
8	07-12-2018 10:45	200,063	179.3	35.99	24.05	2.035	24.82	24.09
9	07-12-2018 10:46	200,064	181	36.07	24.24	2.084	24.91	24.08

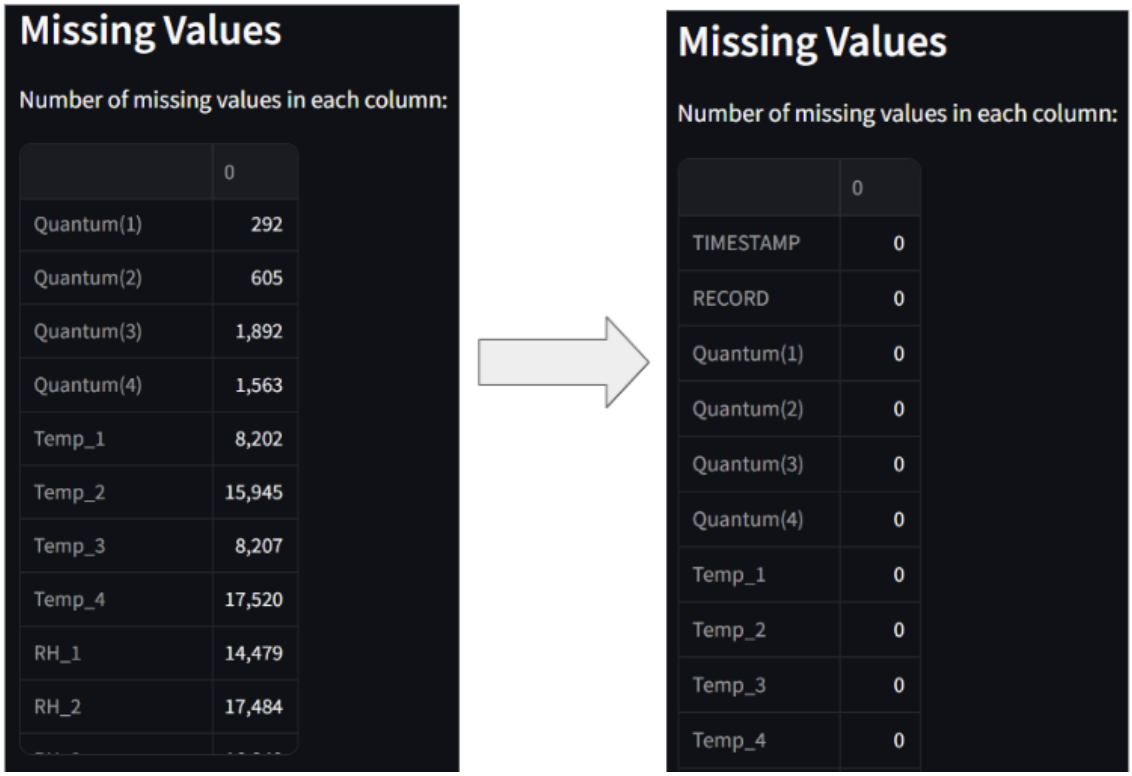
Fonte: Elaboração própria.

Ao inserir dados, eles passam a ficar em *"standby"*, mas ainda não são submetidos de fato. É possível inserir os dados antes de escolher uma campanha para a qual submetê-los, mas a opção de submissão só fica disponível após essa escolha, sendo exibido um aviso caso a campanha não tenha sido selecionada. Um aspecto importante nesse momento é a etapa de visualização, que inclui uma série de informações relevantes para ajudar o usuário a avaliar a qualidade dos dados inseridos antes mesmo do preenchimento do Data Quality Report. As informações fornecidas neste momento são, para cada campo: números de valores ausentes,

total de valores registrados, valores máximos e mínimos, desvio padrão, média e divisão dos quartis.

No figura 11 vemos dois cenários com essa identificação, um com inúmeros dados ausentes, e um em que estes problemas já foram corrigidos.

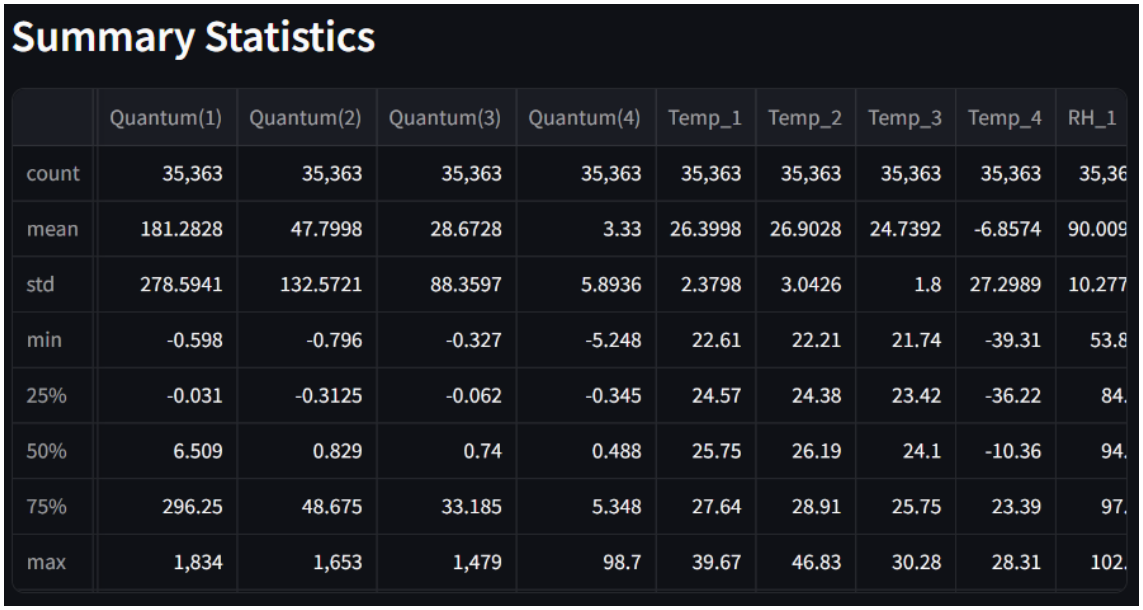
Figura 11 – Dados após serem corrigidos pelo pesquisador



Fonte: Elaboração própria.

Enquanto que na figura 12 vemos um resumo dos dados já corrigidos.

Figura 12 – Resumo de informações relevantes dos dados submetidos



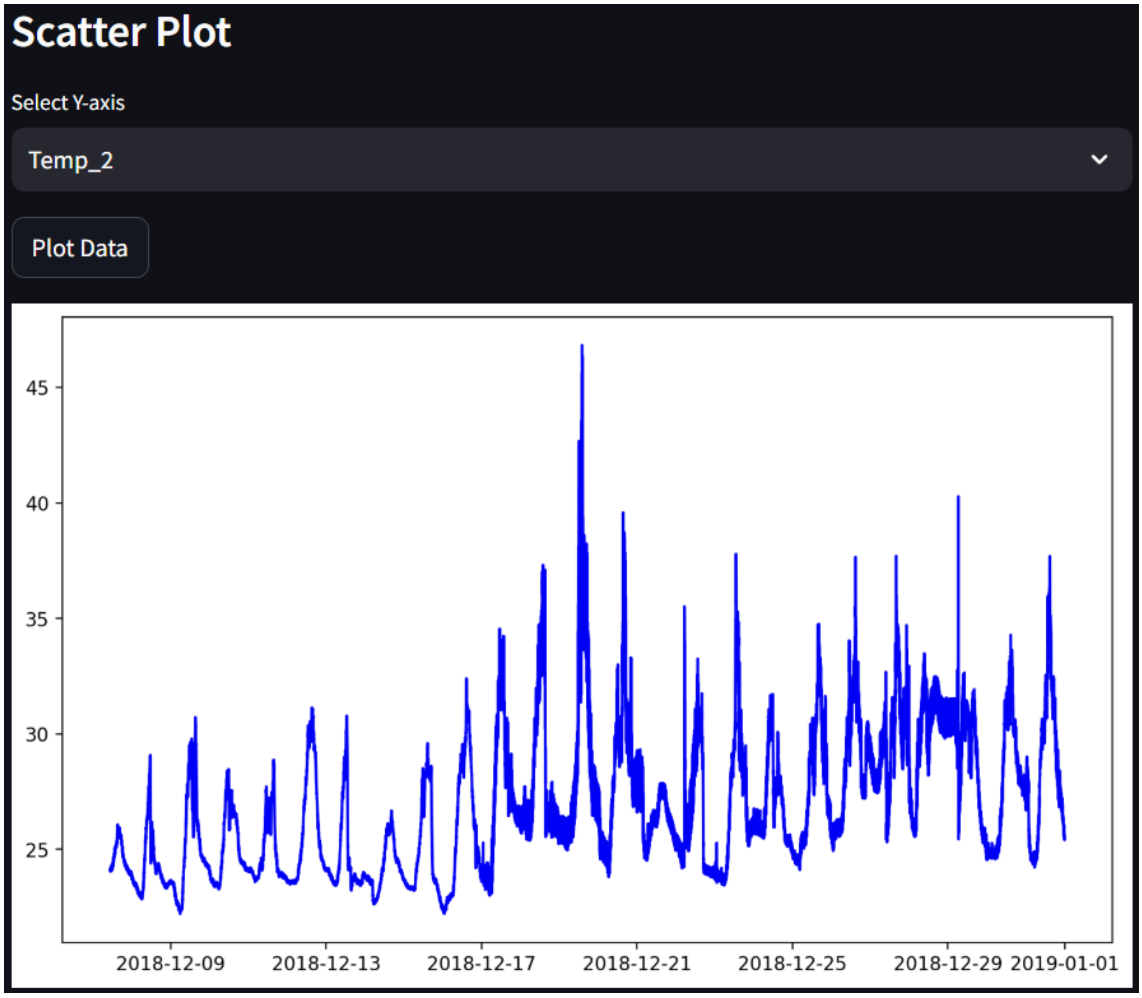
Summary Statistics

	Quantum(1)	Quantum(2)	Quantum(3)	Quantum(4)	Temp_1	Temp_2	Temp_3	Temp_4	RH_1
count	35,363	35,363	35,363	35,363	35,363	35,363	35,363	35,363	35,36
mean	181.2828	47.7998	28.6728	3.33	26.3998	26.9028	24.7392	-6.8574	90.009
std	278.5941	132.5721	88.3597	5.8936	2.3798	3.0426	1.8	27.2989	10.277
min	-0.598	-0.796	-0.327	-5.248	22.61	22.21	21.74	-39.31	53.8
25%	-0.031	-0.3125	-0.062	-0.345	24.57	24.38	23.42	-36.22	84.
50%	6.509	0.829	0.74	0.488	25.75	26.19	24.1	-10.36	94.
75%	296.25	48.675	33.185	5.348	27.64	28.91	25.75	23.39	97.
max	1,834	1,653	1,479	98.7	39.67	46.83	30.28	28.31	102.

Fonte: Elaboração própria.

Por fim, uma última funcionalidade interessante é a capacidade de visualizar cada campo em formato gráfico, facilitando a identificação de erros, como é o caso da figura 13 (há um registro anormal no dia 19).

Figura 13 – Gráfico de dados de temperatura



Fonte: Elaboração própria.

Ao submeter de fato os dados, automaticamente é criada uma nova *collection* para armazená-los, que será nomeada a partir do campaign_id Os dados estão neste ponto preparados para receber sua análise de qualidade. Vale ressaltar que o MongoDB não possui limite para o número de documentos em uma mesma *collection*, ao contrário de ferramentas como o Microsoft Excel, por exemplo, que possui um máximo de 234 células preenchidas. Sendo assim, o único limitante passa a ser a capacidade de armazenamento dos sistemas do próprio administrador, e o sistema se mostra mais robusto ao lidar com aplicações de Big Data.

4.4 Submissão de Data Quality Reports

O Data Quality Report é o elemento fundamental dessa plataforma, e o foco do desenvolvimento até aqui. Após a seleção de uma campanha com dados submetidos, o usuário é redirecionado para uma página em que pode escolher uma medição considerada problemática

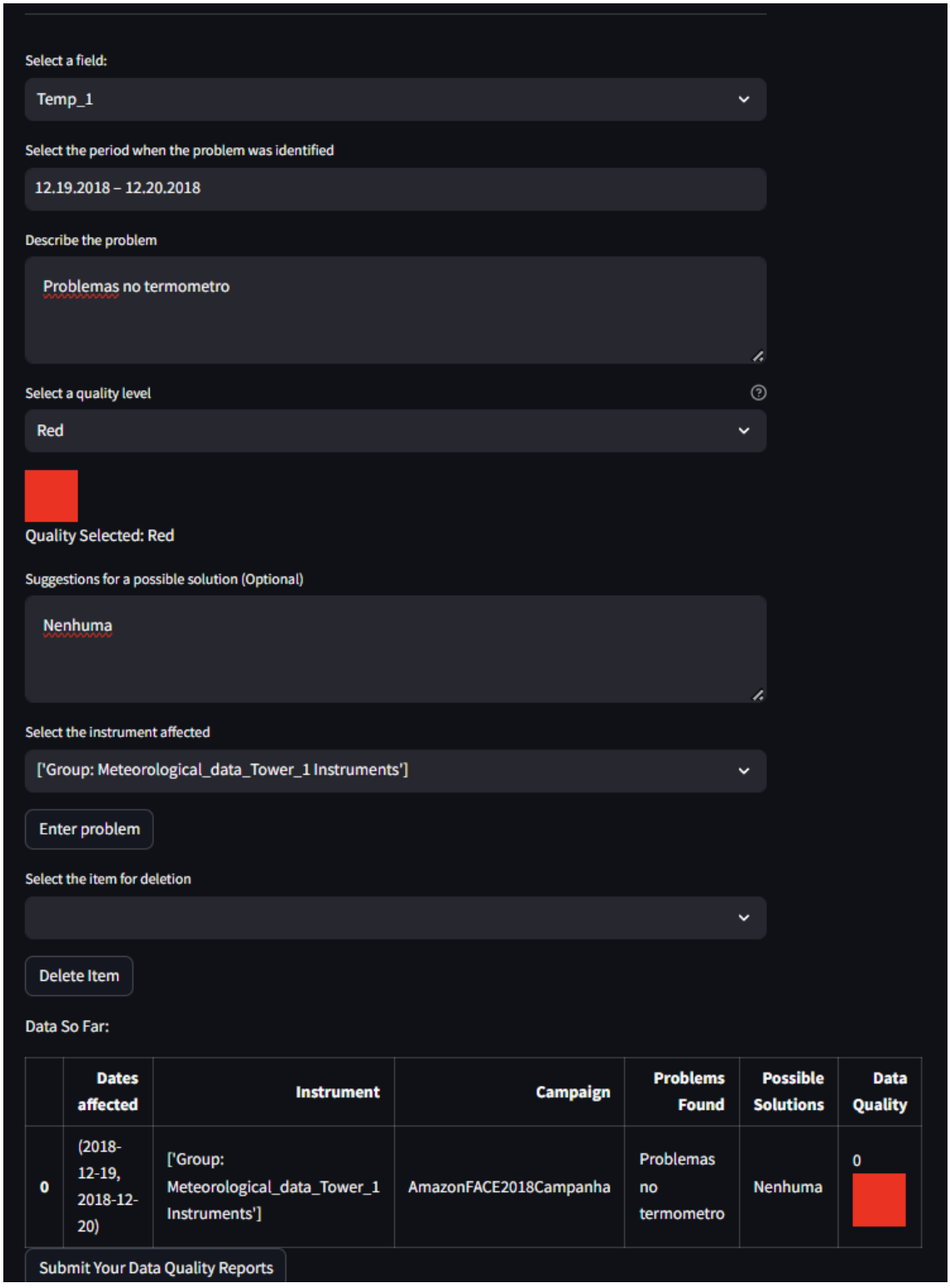
e um período afetado pelo problema. Em seguida, ele deverá descrever o problema e escolher o nível adequado dos dados, que serão explicados a seguir. Ele também deve adicionar os instrumentos afetados, e pode oferecer uma solução caso esta seja possível.

Os 4 níveis a serem escolhidos estão elencados a seguir:

- Vermelho: Dados foram coletados no período definido, mas possuem defeitos que os tornam completamente inutilizáveis. Esse é o caso de falhas consideráveis nos instrumentos. Exemplo: termômetro danificado registrando temperaturas imprevisíveis e inconsistentes com a área de coleta.
- Amarelo: Dados foram coletados no período definido, mas possuem possíveis inconsistências previsíveis e passíveis de correção. Exemplo: termômetro com offset constante de 10°C para todo o período analisado
- Preto: Dados eram esperados para o período avaliado, mas não foram coletados por algum erro. Novamente pode ser um caso de falha considerável que inutilizou a operação de um instrumento. Vale ressaltar que o período ainda deve estar dentro da campanha.
- Roxo: Dados foram coletados no período definido, mas o pesquisador considera suspeitos e possivelmente incorretos, sem contudo ser capaz de confirmar o problema.

Essas informações sobre os níveis estão disponibilizadas por inteiro em um botão de ajuda, ao lado da seleção. Assim, mesmo usuários sem nenhum conhecimento prévio da plataforma são capazes de usá-la facilmente. Na figura 14 podemos ver um exemplo de um Data Quality Report completamente preenchido e submetido.

Figura 14 – Submissão de um Data Quality Report bem sucedida



Fonte: Elaboração própria.

Ao submeter o Data Quality Report, todos os demais dados existentes são considerados corretos. Dessa forma, levando-se em conta que a maior parte dos dados coletados em pesquisa são de fato confiáveis, o preenchimento do Data Quality Report completo fica facilitado. É

importante ressaltar que a responsabilidade pela veracidade do Report realizado recaí sobre o usuário que os submeteu e as instituições associadas a ele.

Também vale pontuar que múltiplos Data Quality Reports podem ser submetidos sobre um único conjunto de dados, tanto para um mesmo período quanto para períodos diferentes (dentro da campanha) possibilitando assim uma caracterização mais completa destes.

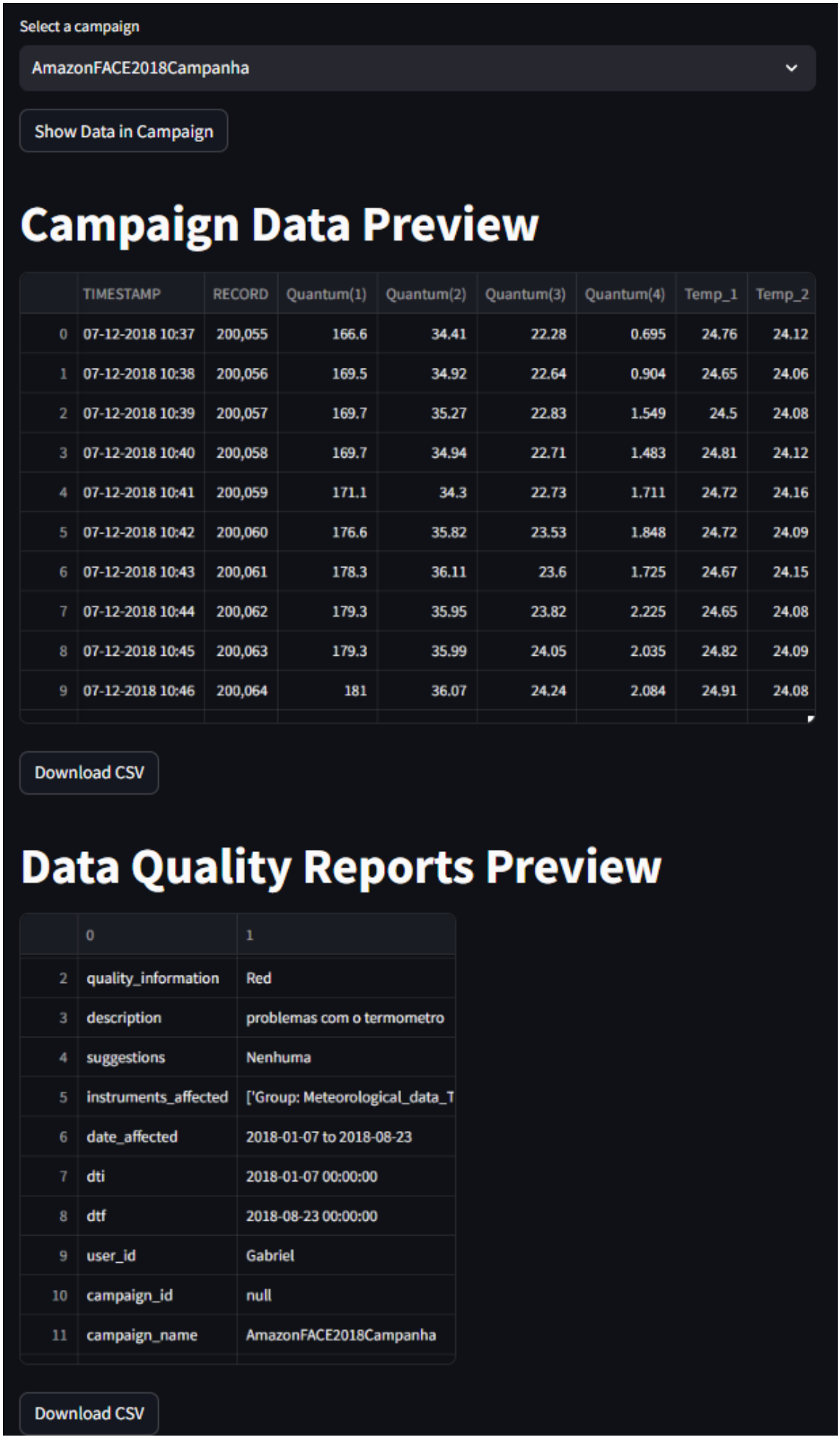
A partir desse momento, os dados ficam disponíveis para qualquer usuários que deseje acessá-lo,

4.5 Recuperação dos dados inseridos

Conforme projetado, qualquer pesquisador interessado é capaz de fazer o download dos dados originais dentro da plataforma, bem como do Data Quality Report associado a ele. Para isso, basta selecionar na página inicial a campanha de seu interesse e os dados associados a ela aos quais ele deseja ter acesso. Ao fazer isso, fica disponível um botão de download e todos os dados. Nota-se que, conforme o esperado, apenas dados que já passaram pela devida avaliação podem ser baixados.

Mantendo-se a consistência de formato, tanto os próprios dados quanto os Data Quality Reports são disponibilizados em formato CSV, e podem ser visualizados antes do download, como visto na figura 15.

Figura 15 – Dados prontos para o download



	TIMESTAMP	RECORD	Quantum(1)	Quantum(2)	Quantum(3)	Quantum(4)	Temp_1	Temp_2
0	07-12-2018 10:37	200,055	166.6	34.41	22.28	0.695	24.76	24.12
1	07-12-2018 10:38	200,056	169.5	34.92	22.64	0.904	24.65	24.06
2	07-12-2018 10:39	200,057	169.7	35.27	22.83	1.549	24.5	24.08
3	07-12-2018 10:40	200,058	169.7	34.94	22.71	1.483	24.81	24.12
4	07-12-2018 10:41	200,059	171.1	34.3	22.73	1.711	24.72	24.16
5	07-12-2018 10:42	200,060	176.6	35.82	23.53	1.848	24.72	24.09
6	07-12-2018 10:43	200,061	178.3	36.11	23.6	1.725	24.67	24.15
7	07-12-2018 10:44	200,062	179.3	35.95	23.82	2.225	24.65	24.08
8	07-12-2018 10:45	200,063	179.3	35.99	24.05	2.035	24.82	24.09
9	07-12-2018 10:46	200,064	181	36.07	24.24	2.084	24.91	24.08

	0	1
2	quality_information	Red
3	description	problemas com o termometro
4	suggestions	Nenhuma
5	instruments_affected	['Group: Meteorological_data_T
6	date_affected	2018-01-07 to 2018-08-23
7	dti	2018-01-07 00:00:00
8	dtf	2018-08-23 00:00:00
9	user_id	Gabriel
10	campaign_id	null
11	campaign_name	AmazonFACE2018Campanha

4.6 Paralelos e diferenças com o sistema ARM

A principal inspiração do presente trabalho foi o sistema de controle de qualidade de dados desenvolvido pelo ARM, como já citado em seções anteriores. Sobre ele, notam-se alguns paralelos e diferenças relevantes, que ajudam a entender melhor tanto as inovações próprias deste projeto, sua capacidade de cumprir com os resultados esperados, e suas limitações diante de um projeto mais maduro e robusto.

Uma das principais diferenças é o formato dos arquivos utilizados, sendo a preferência do ARM, como já também já mencionado, pelo NetCDF.

Tendo esse formato sido desenvolvido para facilitar o controle de metadados e o compartilhamento de dados científicos, ele é muito mais completo nesse sentido, e permite um controle mais sofisticado da qualidade e origem dos dados. Por outro lado, é também um formato complexo, que exige alto nível de conhecimento do pesquisador dos instrumentos utilizados, e portanto dificilmente poderia ser implementado de forma satisfatória dentro do escopo deste trabalho

Outro ponto relevante é a forma e frequência de atualização dos dados cadastrados: enquanto o ARM possui um sistema automatizado que permite a submissão e disponibilização dos dados com poucos minutos de atraso da sua coleta, a plataforma desenvolvida no presente trabalho atualmente depende da submissão manual pelos pesquisadores, de forma que em termos práticos ela se torna incapaz de disponibilizar os dados tão rapidamente.

Por outro lado, como já citado, a plataforma do ARM é inteiramente fechada e de difícil uso por pesquisadores pouco experientes, de forma que a simplicidade das ferramentas desenvolvidas e sua disponibilização livre pode se mostrar mais útil para projetos sem os recursos dessa instituição, bem como mais maleável para customização e adequação a necessidades específicas.

5 CONSIDERAÇÕES FINAIS

Dando continuidade ao desenvolvimento da plataforma de submissão de qualidade de dados, construída utilizando Streamlit e MongoDB, este trabalho concentrou-se nas funcionalidades essenciais para validar os principais conceitos da ferramenta. Considerando o caráter de demonstração e o objetivo de atingir a validação, as funcionalidades de maior valor para o usuário foram priorizadas, resultando em um "MVP"*(Minimum Viable Product)* dentro de uma prova de conceito planejada. Acredita-se que o resultado atingido esteja dentro dos parâmetros esperados.

Ao longo do texto, também foi evidenciado o cumprimento dos requisitos não funcionais de projeto, com o desempenho sendo verificado a partir do funcionamento do sistema em um computador pessoal, capacidades de segurança a partir do uso de bibliotecas com práticas modernas de criptografia e hashing, e facilidade do uso a partir de interfaces e fluxos de interação intuitivos.

Nesse ponto, a única necessidade que não pode ser suficientemente avaliada foi a escalabilidade, visto que a implementação permaneceu contida. Contudo, todas as tecnologias utilizadas foram deliberadamente escolhidas por apresentarem, em suas documentações e propósitos expostos, ampla capacidade de expansão e rápida escalação. Assim, no âmbito da usabilidade do sistema como produto, são vislumbradas implementações em proporções maiores em um futuro próximo.

Para a versão final do produto, há ainda a necessidade de implementar diversas outras funcionalidades, como será visto adiante. Também é necessário se aprofundar no rigor científico da definição de qualidade de dados, e avaliar como as ferramentas desenvolvidas são de fato capazes de a influenciar.

5.1 Perspectivas de Continuidade

Este projeto se encontra inserido dentro de um contexto de atividade de pesquisa e desenvolvimento continuadas, de forma que será

Conforme supracitado, essa fase teve foco na implementação dos primeiros recursos necessários para cumprimento dos requisitos estabelecidos nas primeiras seções e início de atividades de pesquisa mais amplas. Abaixo podem ser vistas algumas oportunidades apresentadas, que servem tanto para melhorar a experiência do usuário quanto para aumentar a utilidade da plataforma, e serão exploradas em um momento futuro.

1. **Submissão de dados em formatos variados:** possibilitar que os dados a serem inseridos na plataforma não estivessem limitados ao formato CSV, incorporando por exemplo NetCDF e dados brutos.

2. **Automatização de parte do processo de criação de Data Quality Reports:** possibilitar que o usuário selecione certas condições para os dados, a partir das quais eles receberiam automaticamente um Data Quality Report. Por exemplo, caso um certo registro excedesse um valor máximo, ele poderia imediatamente ser marcado como incorreto.

3. **Desenvolvimento de uma API para controle da disponibilização de dados:** As ferramentas usadas no desenvolvimento limitam o tamanho máximo dos dados que pode ser transferidos para o usuário, de forma que uma API para melhor controle ao acesso a esses dados, incluindo uma fila enquanto as requisições são processadas, seriam uma boa adição a plataforma

4. **Integração com sistemas acadêmicos**: a plataforma foi construída primariamente para funcionar de forma independente, mas considerando seu foco em pesquisa, seria oportuno uma maior integração com plataformas como ORCID e Lattes;

5. **Separação dos dados por áreas de pesquisa**:atualmente todos os dados submetidos se encontram em uma mesma tabela de pesquisa, mas faz sentido que, conforme o crescimento da plataforma, houvesse um esforço para separa-los em áreas diferentes;

6. **Geração de DQR no formato JSON**: o formato atual em CSV para os Data Quality Reports, embora consistente com os demais formatos usados, pode ser trabalhoso pra manipulação posterior. Assim, esse aspecto pode ser favorecido pela implementação de novos formatos, como é o caso do JSON, que geralmente é mais facilmente adaptável a bibliotecas e frameworks de múltiplas linguagens

7. **Geração de DOI para os dados**: seria interessante que os dados inseridos pudessem receber um DOI (Digital Object Identifier) associado, que representaria uma identificação única dentro dos padrões internacionais, e facilitaria sua busca e reuso.

REFERÊNCIAS

PEPPLER, R. A. et al. The arm data quality program. *Meteorological Monographs*, v. 57, p. 12–1, 2016.

HAZEN, B. T. et al. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, Elsevier, v. 154, p. 72–80, 2014.

BATINI, C. et al. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 41, n. 3, p. 1–52, 2009.

WILKINSON, M. D. et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2016.

BOECKHOUT, M.; ZIELHUIS, G. A.; BREDENOORD, A. L. The fair guiding principles for data stewardship: fair enough? *European journal of human genetics*, Springer International Publishing Cham, v. 26, n. 7, p. 931–936, 2018.

PALANISAMY, G. *ARM Data File Standards Version 1.2*. [S.l.], 2016.

RAMMIG, A.; LAPOLA, D. Amazonface - assessing the response of amazon rainforest functioning to rising atmospheric co2 concentration. *EGU22, the 24th EGU General Assembly, held 23-27 May, 2022 in Vienna, Austria and Online*, 2022. Online at https://egu22.eu/, id.EGU22-9067. Disponível em: ⟨https://egu22.eu/⟩.

GIEBLER, C. et al. A Zone Reference Model for Enterprise-Grade Data Lake Management. In: *Proceedings of the 24th IEEE Enterprise Computing Conference (EDOC 2020)*. [S.l.: s.n.], 2020.

BLINOWSKI, G.; OJDOWSKA, A.; PRZYBYŁEK, A. Monolithic vs. microservice architecture: A performance and scalability evaluation. *IEEE Access*, v. 10, p. 20357–20374, 2022.

GONCALVES, G. G. P. *DataQualityDataMap*. 2023. ⟨https://github.com/GandraOProprio/DataQualityDataMap⟩. Acesso em 12 de dezembro de 2023.