IV Simpósio Brasileiro de Infraestrutura de Dados Espaciais - SBIDE
INDE Infraestrutura Nacional de Dados Espaciais
IDEs Inteligentes para governança e políticas públicas
15 a 18 de outubro de 2024

IS_AGRO, MÓDULO DIGITAL: O USO DA ARQUITETURA MEDALLION COMO BASE PARA AUTOMAÇÃO DE ROTINAS DE EXECUÇÃO DE PIPELINES

Carlos Eduardo da Silva Sacramento – Embrapa Solos
Carlos Eduardo Miranda Mota – SGB/CPRM
Edgar Shinzato – SGB/CPRM
Pedro Luiz de Freitas – Embrapa Solos

APRESENTAÇÃO

O projeto IS_Agro (Figura 1) é uma iniciativa voltada à avaliação crítica e à subsequente adaptação de metodologias concebidas em fóruns globais (destacadamente FAO e OCDE), com vistas à sua aplicação no contexto nacional a partir da elaboração de novas métricas e indicadores agro-socioambientais (IASs) que almejam fornecer uma representação mais precisa e autêntica do panorama agropecuário em território nacional, monitorando e avaliando o desempenho agropecuário relacionado aos aspectos sociais, econômicos e ambientais, servindo "para avaliar a performance da agropecuária quanto ao seu desempenho ambiental, social e econômico, fornecendo dados e informações comparativos entre as entidades federativas ou países, dentre diversas outras aplicações" [1]. A concepção deste projeto como uma plataforma digital vinculada ao Observatório da Agropecuária Brasileira almeja publicar indicadores e parâmetros oriundos de dados técnico-científicos embasados, aptos a avaliar o efetivo desempenho do setor agropecuário nacional a nível municipal ou estadual, contribuindo com as políticas setoriais e os processos de planejamento e gestão que visam à edificação de uma agropecuária sustentável e ao correto posicionamento do país no cenário internacional. Assim, o objetivo geral é o desenvolvimento de um ambiente inteligente que administre a automação de *pipelines* dos IASs em um ambiente de armazenamento de dados organizado sob a arquitetura *medallion* (arquitetura de medalhas) como base do painel de dados (*dashboards*) da publicação dos indicadores.

IS AGRO Indicadores Agro-socioambientais — Indicadores agro-socioambientais do Brasil: inteligência estratégica para a sustentabilidade da agropecuária nacional

Figura 1: Logotipo do projeto IS_Agro.

MATERIAIS E MÉTODOS

Neste projeto, os IASs são desenvolvidos por diferentes equipes especializadas nas temáticas propostas, cujos trabalhos são previamente aprovados e publicados no cenário científico. Para automação das coletas dos dados, alocação, cálculos e constantes atualizações dos IASs há a equipe do chamado Módulo Digital, que desenvolve soluções para cada indicador, transformando-os em algoritmos digitais. São coletados dados cadastrais estruturados, semi-estruturados e não estruturados guardados em um *data lakehouse*, exigindo uma grande organização dentro do repositório para que os dados estejam sempre disponíveis e que tenham fácil acesso. Decidiu-se implantar a arquitetura *medallion*, enquanto para gestão e automação dos pipelines foi utilizado uma plataforma de código aberto. A arquitetura *medallion* é a estruturação sequencial de armazenamento de dados que visa organizar logicamente os dados do *lakehouse*, objetivando a melhora de forma incremental e progressiva da estrutura e da qualidade dos dados à medida que fluem pelas três camadas da arquitetura [4]. Os termos bronze (dados brutos da fonte), prata (transformação e validação dos dados) e ouro (dados refinados e enriquecidos para uso de projetos) descrevem a qualidade dos dados durante o processo [5] (Figura 2).

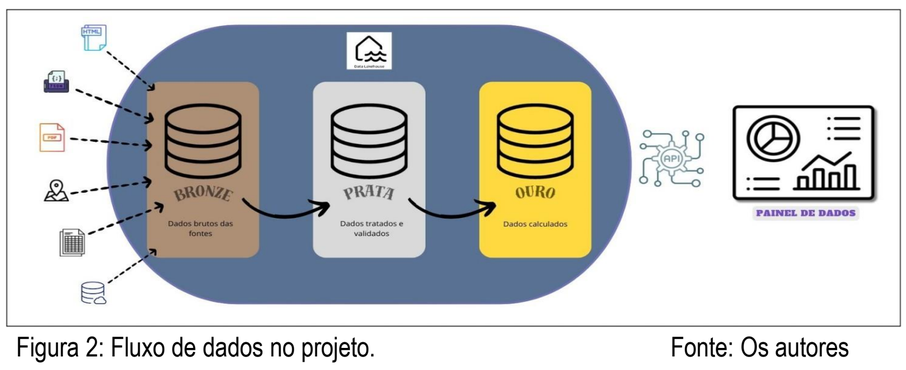

Figura 2: Fluxo de dados no projeto. Fonte: Os autores

Os códigos foram desenvolvidos em python no ambiente do Visual Studio Code. O gerenciamento dos *pipelines* é executado pelo Apache Airflow (Figura 3), plataforma de código aberto para desenvolvimento, agendamento e monitoramento de fluxos de trabalho orientados em lote sob estrutura da linguagem de programação python que permite criar fluxos de trabalho conectados a praticamente qualquer tecnologia [6]. O ambiente de execução do Airflow foi estruturado em Docker e a imagem desenvolvida está disponibilizada no GitHub, permitindo mobilidade e flexibilidade em sua instalação.

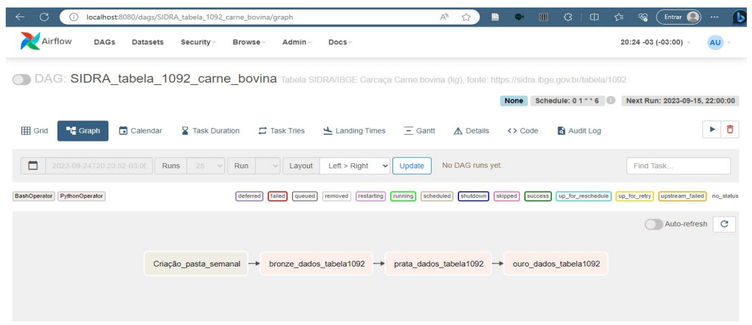

Figura 3: Tela do Airflow com um pipeline. Fonte: Os autores

A previsão inicial da periodicidade de execução automática das rotinas é de uma vez por mês. A coleta dos dados brutos se dá em forma de *download* com a manutenção do seu formato original. Junto é gravado um hash de cada arquivo para, em caso de mudança, indicar que houve atualização dos dados e realizar novo *download*. Esses dados são higienizados e tratados conforme necessidade. Ao final da fase prata, uma estrutura tabular verticalizada será salva no data lakehouse como .parquet, formato de código aberto de armazenamento colunar projetado para armazenamento de alta compactação e recuperação eficiente de dados, fornecendo desempenho aprimorado para lidar com dados complexos em massa [7]. Os .parquet salvos ficam disponíveis para uso na camada ouro com cardinalidade um para muitos. Nesta última fase da arquitetura são aplicados cálculos diretos das tabelas tratadas. Após, esses dados de nível ouro são exportados um banco de dados do projeto no PostgreSQL, prontos para uso por uma API desenvolvida internamente que permita o fornecimento dos dados para o painel de dados a ser desenvolvido (Figura 4) e publicado para a sociedade (ou seja, aberto a todos) a partir do sítio do projeto na internet.

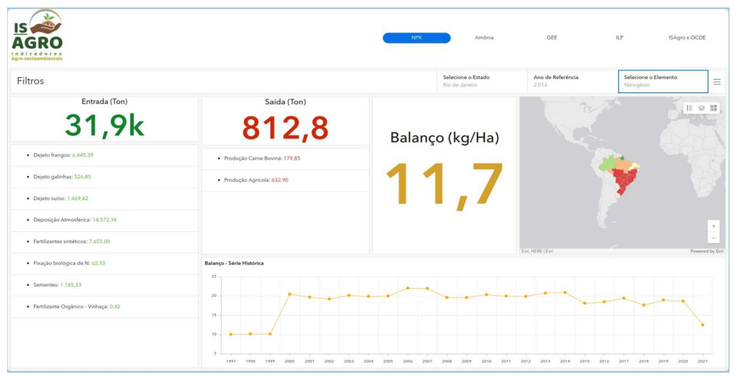

Figura 4: Painel de dados visualizando um dos indicadores.

RESULTADOS E DISCUSSÕES

Há indicadores que não seguem uma estrutura linear de seu desenvolvimento, ou seja, necessitam do cruzamento de tabelas. A arquitetura *medallion* é flexível quanto ao uso das tabelas, permitiu o uso delas (em níveis prata e ouro) para mais de um indicador, otimizando os processos. Além disso funciona incrementalmente, baixando e processando apenas os dados que sofreram alguma modificação em sua origem e recalculando todos os pipelines relacionados. Baixar os arquivos brutos garantiu que os dados pudessem ser recriados sempre que houve necessidade sem precisar de nova conexão, que depende do acesso às fontes estarem disponíveis. E, ao não excluir os antigos dados baixados quando novos são baixados, criou-se um histórico que possibilita uma investigação caso exista necessidade. Por fim, a compreensão do fluxo de dados foi amplamente facilitada para outras pessoas devido a simples e bem definida divisão de camadas da arquitetura.

CONSIDERAÇÕES FINAIS

Este trabalho não objetivou detalhar cada indicador e suas fontes e sim destacar a arquitetura de dados e suas vantagens. O modelo de desenvolvimento de dados pelo projeto veio sendo ajustado e corrigido com o aumento do volume de dados. A utilização da arquitetura medallion permite esses tipos de ajustes sem que seja preciso refazer do início. Flexível, está pronto para receber novos indicadores desenvolvidos por outras equipes, assim como o desenvolvimento do painel de dados para a publicação dos dados.

REFERÊNCIAS

[1] EMBRAPA SOLOS (org.). INDICADORES agro-socioambientais do Brasil: inteligência estratégica para a sustentabilidade da agropecuária nacional. Rio de Janeiro: Embrapa Solos, jul 2023.
[4] ARQUITETURA medallion. In: DATABRICKS (E.U.A.). Glossário. 2024?.
[5] SKAYA, I. et al. O que é arquitetura medallion do Lakehouse?. In: MICROSOFT (E.U.A.) (org.). Microsoft Learn: Azure Databricks documentation. 1 mar. 2024.
[6] WHAT is Airflow™?. In: APACHE AIRFLOW. Documentation: Apache Airflow. 2023.
[7] OVERVIEW. In: APACHE PARQUET (org.). Documentation. 24 abr. 2022.