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

CARLOS EDUARDO DA SILVA SACRAMENTO [1]
CARLOS EDUARDO MIRANDA MOTA [2]
EDGAR SHINZATO [3]
PEDRO LUIZ DE FREITAS [4]

[1]EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA EMBRAPA SOLOS, RIO DE JANEIRO - RJ
EDUSACRAMENTO@AOL.COM

[2]SERVIÇO GEOLÓGICO DO BRASIL DIVISÃO DE GEOPROCESSAMENTO
DEPARTAMENTO DE INFORMAÇÕES INSTITUCIONAIS DIRETORIA DE ESTRUTURA GEOCIENTÍFICA, RIO DE JANEIRO - RJ
CARLOS.MOTA@SGB.GOV.BR

[3]SERVIÇO GEOLÓGICO DO BRASIL DEPARTAMENTO DE INFORMAÇÕES INSTITUCIONAIS
DIRETORIA DE ESTRUTURA GEOCIENTÍFICA, RIO DE JANEIRO - RJ
EDGAR.SHINZATO@SGB.GOV.BR

[4]EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA EMBRAPA SOLOS, RIO DE JANEIRO - RJ
PEDRO.FREITAS@EMBRAPA.BR

O projeto IS_Agro é uma iniciativa voltada à avaliação crítica e à subsequente adaptação de metodologias concebidas em fóruns globais, 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. Os IASs são medidas utilizadas para monitorar e avaliar o desempenho agropecuário relacionado aos aspectos sociais, econômicos e ambientais, assim tendo grande importância na orientação de estratégias políticas e práticas agrícolas mais sustentáveis, seja pelo ente público ou privado, 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]. 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* (arquitetura de medalhas), que consiste na alocação de dados em três camadas com diferentes finalidades, enquanto para gestão e automação dos *pipelines* foi utilizado uma plataforma de código aberto.

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 automatize e administre os *pipelines* dos IASs em um ambiente de organização de armazenamento de dados baseado na arquitetura *medallion* (Figura 1) para ser base do painel de dados da publicação dos indicadores.

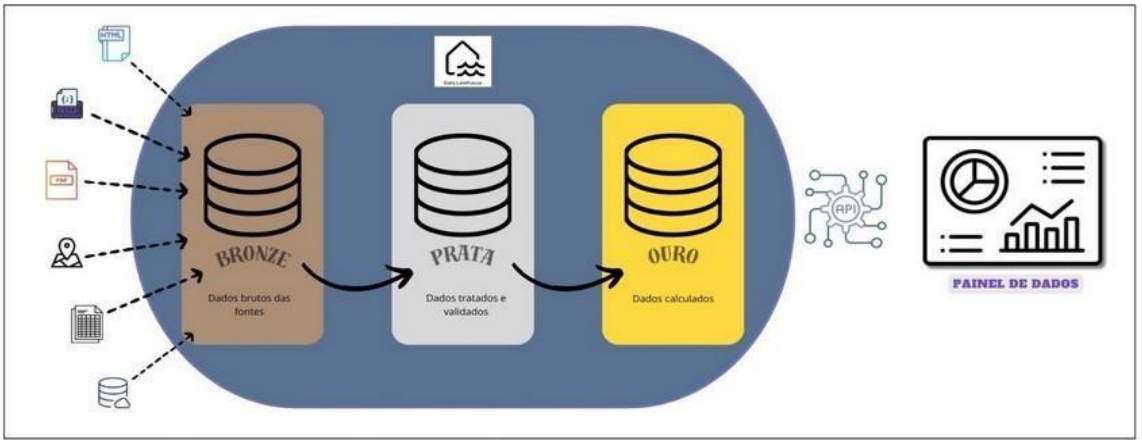

Figura 1. Fluxo de dados na arquitetura medallion no projeto

Um *pipeline* de dados é uma sucessão de fases conectadas que permitem a coleta, o armazenamento, a modificação, a análise e a representação de dados, tendo como propósito adquirir conhecimentos significativos e embasar escolhas esclarecidas [2]. Já um *data lakehouse*, destino dos *pipelines* do projeto, é "como uma plataforma de dados moderna construída a partir de uma combinação de um *data lake* e um *data warehouse"* [3], utilizando "o armazenamento flexível de dados não estruturados de um *data lake* e os recursos e ferramentas de gerenciamento de *data warehouses* e, em seguida, implementá-los estrategicamente juntos como um sistema maior" [3]. 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]. O gerenciamento dos *pipelines* é executado pelo Apache Airflow (versão 2.5.3), 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, plataforma de código aberto que permite a criação e a administração de contêineres como máquinas virtuais modulares que contém o essencial para sua execução. A imagem desenvolvida está disponibilizada no GitHub.

A ser confirmada, a periodicidade de execução das rotinas é de uma vez por mês. A coleta dos dados brutos se dá em forma de *download* e manutenção do seu formato original, sendo 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 estará como geocodigo (inteiro, código IBGE dos municípios ou UFs), data (marca temporal, ISO 8601), fonte (texto) e valor (ponto flutuante, número real) e 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 no lago de dados ficam disponíveis para uso na camada ouro com cardinalidade um para muitos. Nesta última fase da arquitetura são realizados os cálculos necessários para cada fonte dos indicadores, havendo fontes que não necessitam de cálculos. A finalização é com a exportação dos dados ouro para tabelas em um banco de dados do projeto no PostgreSQL, estando prontos para uso por uma API desenvolvida internamente que permita o fornecimento dos dados para o painel de dados a ser desenvolvido (por outra equipe) e publicado para a sociedade a partir do sítio do projeto na internet (Figura 2).

Figura 2. Painel de dados de um dos indicadores agro socioambientais do IS_Agro.

Este modelo veio sendo ajustado e corrigido ao longo do desenvolvimento do projeto no Módulo Digital. Flexível, hoje é considerado pronto para receber qualquer indicador desenvolvido pelas outras equipes, assim como o desenvolvimento do painel de dados para publicação para uso da sociedade.

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. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/doc/1154982/1/Indicadore s-agro-socioambientais-do-Brasil-2023.pdf>. Acesso em: 17 ago. 2023.

[2] CALANCA, P. O que é um pipeline de dados?. In: ESCOLA DE DATA SCIENCE, ALURA (Brasil). Data Science. Brasil, 26 jul. 2023. Disponível em: <https://www.alura.com.br/artigos/o-que-pipeline-dados>. Acesso em: 21 ago. 2023.

[3] ORACLE CLOUD INFRASTRUCTURE (Brasil). O que é um Data Lakehouse?. In: Big Data. 2023?. Disponível em: <https://www.oracle.com/br/big-data/what-is-data-lakehouse/>. Acesso em: 2 set. 2023.

[4] ARQUITETURA medallion. In: DATABRICKS (E.U.A.). Glossário. 2024?. Disponível em: <https://www.databricks.com/br/glossary/medallion- architecture>. Acesso em: 11 jul. 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. Disponível em: <https://learn.microsoft.com/pt-br/azure/databricks/lakehouse/medallion>. Acesso em: 11 jul. 2024.

[6] WHAT is Airflow™?. In: APACHE AIRFLOW. Documentation: Apache Airflow. 2023. <Disponível em: https://airflow.apache.org/docs/apache- airflow/stable/index.html>. Acesso em: 22 set. 2023.

[7] OVERVIEW. In: APACHE PARQUET (org.). Documentation. 24 abr. 2022. Disponível em: <https://parquet.apache.org/docs/overview/>. Acesso em: 21 set. 2023.