BOCA
BOLETIM DE CONJUNTURA

PROTÓTIPO DE UM SISTEMA DE RECOMENDAÇÃO DE DADOS AGRÍCOLAS BASEADO EM UMA PLATAFORMA DE DADOS ESCALÁVEL

PROTOTYPE OF AN AGRICULTURAL DATA RECOMMENDATION SYSTEM BASED ON A SCALABLE DATA PLATFORM

PROTOTIPO DE UN SISTEMA DE RECOMENDACIÓN DE DATOS AGRÍCOLAS BASADO EN UNA PLATAFORMA DE DATOS ESCALABLE

10.56238/bocav25n74-020

Rafael Ignaulin
Graduado em Ciência da Computação
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: rafael.ignaulin@unochapeco.edu.br

Sandro Silva de Oliveira
Doutor em Tecnologia e Gestão da Inovação
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: silva@unochapeco.edu.br

Cristiano Reschke Lajús
Doutor em Agronomia
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: clajus@unochapeco.edu.br

Ariel Gustavo Zuquello
Doutor em Ciência e Engenharia dos Materiais
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: Ariel.zuquello@unochapeco.edu.br

Éttore Guilherme Poletto Diel
Graduando em Agronomia
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: ettorepoletto77@gmail.com

Fábio José Busnello
Doutor em Agronomia
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: fbusnello@yahoo.com.br

Magdalena Reschke Lajús Travi
Doutora em Agronomia
Instituição: Universidade Comunitária da Região de Chapecó (UNOCHAPECÓ)
E-mail: magdalena@unochapeco.edu.br

Mauricio Bedin
Graduado em Engenharia Mecânica
Instituição: Unidade Central de Educação Faem Faculdade (UCEFF)
E-mail: mauriciobedin-@hotmail.com

RESUMO
Este projeto visa desenvolver uma plataforma de dados robusta para análises exploratórias na área técnica da agronomia. Foram utilizados conceitos de arquitetura de aplicações e computação em nuvem voltados ao processamento de big data. Diferentes tipos de dados foram centralizados em uma plataforma escalável e segura, utilizando serviços da AWS, reconhecidos por sua confiabilidade e alta disponibilidade. A solução incorporou ferramentas open-source como Apache Spark e Apache Airflow, responsáveis pelo processamento e orquestração distribuída dos pipelines de dados. As principais fontes de dados incluíram informações meteorológicas do Instituto Nacional de Meteorologia (INMET) e dados agrícolas do Ministério da Agricultura e Pecuária (MAPA). A estrutura proposta permitiu o processamento detalhado dessas informações, possibilitando a geração de insights relevantes. Foram realizadas análises que investigaram a influência da soma térmica e da pluviometria na produtividade de diferentes híbridos de milho, segmentadas por localização e período. Como produto final, foi desenvolvido um dashboard interativo e intuitivo, permitindo que agrônomos visualizem dados históricos e realizem projeções futuras com base nas informações processadas. Conclui-se que a plataforma oferece uma base resiliente, escalável e eficiente para o tratamento de grandes volumes de dados, com grande potencial para apoiar decisões técnicas mais precisas e fundamentadas no setor agrícola.

Palavras-chave: Produção Agrícola. Tecnologia. Análise de Dados.

ABSTRACT
This project aims to develop a robust data platform for exploratory analyses in the technical field of agronomy. Concepts of application architecture and cloud computing focused on big data processing were applied. Different types of data were centralized in a scalable and secure platform using AWS services, known for their reliability and high availability. The solution incorporated open-source tools such as Apache Spark and Apache Airflow, responsible for distributed processing and orchestration of data pipelines. The main data sources included meteorological information from the National Institute of Meteorology (INMET) and agricultural data from the Ministry of Agriculture and Livestock (MAPA). The proposed structure enabled detailed processing of this information, allowing for the generation of relevant insights. Analyses were conducted to investigate the influence of thermal sum and rainfall accumulation on the productivity of different corn hybrids, segmented by location and time period. As a final product, an interactive and intuitive dashboard was developed, allowing agronomists to visualize historical data and make future projections based on the processed information. It is concluded that the platform offers a resilient, scalable, and efficient foundation for handling large volumes of data, with great potential to support more precise and well-informed technical decisions in the agricultural sector.

Keywords: Agricultural Production. Technology. Data Analysis.

RESUMEN
Este proyecto busca desarrollar una plataforma de datos robusta para análisis exploratorios en el campo técnico de la agronomía. Se utilizaron conceptos de arquitectura de aplicación y computación en la nube enfocados en el procesamiento de big data. Se centralizaron diferentes tipos de datos en una plataforma escalable y segura utilizando servicios de AWS, reconocidos por su confiabilidad y alta disponibilidad. La solución incorporó herramientas de código abierto como Apache Spark y Apache Airflow, responsables del procesamiento distribuido y la orquestación de los flujos de datos. Las principales fuentes de datos incluyeron información meteorológica del Instituto Nacional de Meteorología (INMET) y datos agrícolas del Ministerio de Agricultura y Ganadería (MAPA). La estructura propuesta permitió un procesamiento detallado de esta información, lo que permitió la generación de información relevante. Se realizaron análisis para investigar la influencia de la suma térmica y la precipitación en la productividad de diferentes híbridos de maíz, segmentados por ubicación y período. Como producto final, se desarrolló un tablero interactivo e intuitivo que permite a los agrónomos visualizar datos históricos y realizar proyecciones futuras basadas en la información procesada. Se concluye que la plataforma ofrece una base resiliente, escalable y eficiente para el procesamiento de grandes volúmenes de datos, con gran potencial para respaldar decisiones técnicas más precisas e informadas en el sector agrícola.

Palabras clave: Producción Agrícola. Tecnología. Análisis de Datos.

1 INTRODUÇÃO

Na era digital, a sociedade está imersa em um ambiente altamente conectado, no qual bilhões de informações são geradas e compartilhadas diariamente em plataformas como redes sociais e dispositivos inteligentes. Essa enorme quantidade de dados em tempo real representa um grande desafio para indivíduos e organizações, pois exige soluções eficazes para seu gerenciamento e análise. Considerados o recurso mais valioso do século XXI, os dados permitem identificar padrões e tendências relevantes, oferecendo vantagens competitivas significativas a quem consegue utilizá-los de forma estratégica, especialmente no entendimento de comportamentos e dinâmicas de mercado.

Nesse contexto, surge o conceito de Big Data, que engloba um conjunto de tecnologias e métodos voltados para a coleta, o armazenamento, o processamento e a análise de grandes volumes de dados, com o objetivo de extrair insights relevantes para a tomada de decisões. Como destaca Kleppmann (2017), empresas como Google, Facebook e Amazon processam quantidades massivas de dados para responder rapidamente às mudanças de mercado, testar hipóteses e gerar novos conhecimentos. Assim, este trabalho propõe a criação de uma plataforma que simplifique o uso de dados, facilitando sua coleta e processamento, a fim de apoiar decisões mais assertivas por meio da geração de insights.

Além disso, a crescente disponibilidade de dados apresenta oportunidades valiosas para a sociedade, especialmente no setor agrícola. No entanto, um dos principais desafios enfrentados pela tecnologia atual é a correta coleta, processamento e análise desses dados. Conceitos de Big Data surgem justamente para enfrentar esse cenário, buscando soluções para a indisponibilidade e a descentralização das informações, que muitas vezes se encontram dispersas em diferentes fontes. Essa fragmentação dificulta a construção de modelos de processamento confiáveis e, consequentemente, compromete a qualidade das análises e das decisões baseadas nesses dados.

Assim, a agricultura e a produção de alimentos, em especial, possuem grande potencial de transformação por meio do uso inteligente de dados. Nos últimos anos, diversos estudos têm se dedicado a identificar gargalos do setor e aumentar a produtividade das lavouras com base na análise integrada de informações. Nesse contexto, torna-se essencial o desenvolvimento de uma plataforma que unifique dados de múltiplas fontes, viabilize seu processamento e os disponibilize de forma acessível para os agentes do setor, como agrônomos e produtores. Essa centralização permitirá gerar insights mais precisos e resolver problemas com base em evidências concretas, promovendo uma agricultura mais eficiente e sustentável.

Em suma, o objetivo geral deste trabalho é construir um protótipo de plataforma de dados voltada para aplicações e análises relacionadas ao manejo de plantas e à agricultura, com foco na melhoria da produtividade e na tomada de decisão baseada em dados. Para isso, pretende-se

implementar um pipeline de dados robusto, capaz de lidar com grandes volumes de informações no contexto do Big Data. A plataforma será desenvolvida utilizando conceitos de computação em nuvem, garantindo escalabilidade, disponibilidade e acessibilidade global.

Além disso, o projeto buscará integrar sistemas e aplicações já existentes na área da Agronomia, promovendo a centralização e o controle das diversas fontes de dados atualmente descentralizadas. A partir dessa unificação, serão realizadas análises específicas que permitirão extrair insights relevantes para o setor agrícola. Por fim, a plataforma será desenvolvida com foco na usabilidade, disponibilizando os dados de forma simples, intuitiva e funcional, facilitando a tomada de decisão por parte de agrônomos, produtores e demais profissionais envolvidos na cadeia produtiva.

Nesse sentido, a crescente geração de dados na Agronomia, oriundos de pesquisas de campo, monitoramentos climáticos e análises do desenvolvimento vegetal, representa uma oportunidade estratégica para transformar as práticas agrícolas por meio de decisões baseadas em evidências. No entanto, a fragmentação dessas informações em diferentes fontes e localidades compromete o aproveitamento de seu potencial, dificultando análises integradas e a identificação de padrões relevantes. Essa limitação impede que agrônomos, pesquisadores e produtores tenham uma visão ampla e confiável do cenário agrícola, comprometendo o avanço das práticas modernas e a maximização da produtividade.

Diante desse contexto, justifica-se o desenvolvimento de uma plataforma de dados baseada em Big Data, com foco na centralização, integração e análise de informações agronômicas. Ao criar um ambiente robusto e acessível para consolidar dados de manejo, clima e desenvolvimento das plantas, a proposta busca suprir uma lacuna crítica no setor, promovendo análises mais precisas e decisões mais embasadas. Essa solução tecnológica poderá impulsionar tanto a pesquisa científica quanto a eficiência operacional no campo, contribuindo para o aumento da produção de alimentos, a redução de custos e o avanço sustentável da agricultura.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 AGRICULTURA, PRODUÇÃO DE ALIMENTOS E MANEJO DE PLANTAS

Nos últimos 50 anos, o crescimento exponencial da população mundial transformou profundamente os contextos socioeconômico e ambiental, triplicando a demanda por alimentos e exigindo a modernização da agricultura, inclusive em ambientes adversos (KAMILARIS; KARTAKOULLIS; PRENAFETA-BOLDÚ, 2017). A escassez de alimentos, especialmente em sociedades subdesenvolvidas, tornou-se um dos maiores desafios atuais e a tecnologia com destaque para a biotecnologia tem se mostrado essencial para ampliar a produção e reduzir custos (EDWARDS, 2020). Essa modernização se intensificou com a adoção de sementes mais produtivas, sensores remotos, computação em nuvem e

Internet das Coisas, consolidando o conceito de Agricultura Inteligente (KAMILARIS; KARTAKOULLIS; PRENAFETA-BOLDÚ, 2017).

De acordo com Edwards (2020), o conceito de Agricultura Sustentável surgiu nos anos 1980 como resposta à industrialização desenfreada da agricultura, que até então priorizava apenas o aumento da produção para atender à crescente demanda populacional. Esse novo modelo propõe uma produção de alimentos que preserve o meio ambiente, valorize o produtor rural e assegure a viabilidade econômica a longo prazo, considerando a agricultura como um processo biológico. Na prática, busca-se imitar as características naturais dos ecossistemas, reduzindo o uso de insumos químicos e adotando práticas de manejo sustentáveis (KAMILARIS; KARTAKOULLIS; PRENAFETA-BOLDÚ, 2017). Além disso, há uma forte preocupação com a conservação dos recursos naturais, a segurança alimentar, a qualidade de vida dos trabalhadores e a estabilidade ecológica, exigindo um conhecimento aprofundado dos sistemas agrários e o uso de tecnologias como Big Data para monitoramento e tomada de decisão (EDWARDS, 2020).

2.1.1 Agricultura de Precisão

Segundo Kamilaris, Kartakoullis e Prenafeta-Boldú (2017), a Agricultura de Precisão tem ganhado destaque na agricultura moderna por seu potencial de aumentar a produtividade das culturas enquanto reduz o uso de recursos, por meio de tecnologias como GPS e modelos de crescimento que permitem monitorar em tempo real a variabilidade do solo e das plantações. Essa abordagem, voltada para a sustentabilidade e o controle técnico do plantio, já apresenta avanços significativos em pesquisas e, na prática, tem gerado resultados positivos para muitos produtores, superando inclusive o Retorno sobre o Investimento (ROI) em diversas aplicações (AKHTER; SOFI, 2022).

Assim, segundo Akhter e Sofi (2022), as principais atividades relacionadas à Agricultura de Precisão envolvem o planejamento e seleção do solo, com foco em garantir fertilidade e ampliar áreas cultiváveis. A identificação precoce de insetos e pragas, essenciais para evitar prejuízos significativos. O monitoramento do desempenho das plantas com o uso de tecnologias como a Internet das Coisas (IoT), que fornece dados estatísticos valiosos. O acompanhamento dos aspectos físicos do ambiente, como solo e ar. E, o controle preciso da irrigação e da nutrição das plantas, contribuindo para uma gestão mais eficiente e sustentável da produção agrícola.

2.1.2 Manejo de plantas

O manejo das plantações é essencial para aumentar a produção de alimentos e reduzir os custos com recursos, sendo um componente indispensável da agricultura sustentável. O manejo convencional, ou químico, baseia-se no uso intensivo de fertilizantes sintéticos e pesticidas para elevar a produtividade e controlar pragas e doenças, priorizando a eficiência produtiva em detrimento das questões ambientais e

de saúde. Contudo, esse modelo pode causar diversos impactos negativos, como contaminação do solo e da água, perda de nutrientes e resistência de pragas, o que tem motivado a busca por alternativas mais sustentáveis (DELGADO et al., 2019).

Entre essas alternativas, Delgado et al., (2019) citam que se destaca o manejo biológico, que substitui os pesticidas químicos pelo uso de organismos vivos, como predadores naturais e patógenos, para controlar pragas de maneira ecológica e com menor risco à saúde humana. Essa prática é mais sustentável e menos sujeita à resistência de pragas, sendo alvo de intensas pesquisas na biotecnologia para o desenvolvimento de soluções mais eficazes e acessíveis. Além disso, o manejo agro-biológico surge como uma abordagem híbrida, que combina técnicas do manejo biológico e químico, como o uso de pó de rocha para otimizar a produtividade com menor impacto ambiental.

2.2 DADOS E TECNOLOGIA

2.2.1 Data Analytics

Com o crescimento exponencial da geração de dados impulsionado pelas tecnologias de informação e comunicação, o *Data Analytics* tornou-se uma ferramenta essencial para extrair insights relevantes a partir de grandes volumes de dados heterogêneos, como textos, imagens, áudios e vídeos, auxiliando na tomada de decisões estratégicas e informadas. Na agricultura, seu uso tem contribuído significativamente para o aumento da produtividade e a redução do esforço manual, especialmente quando integrado a tecnologias como a Internet das Coisas (IoT) e o aprendizado de máquina (AKHTER; SOFI, 2022). A coleta de dados por sensores IoT em atividades como o manejo de plantas, por exemplo, pode ser processada por sistemas de *Big Data Analytics*, gerando informações valiosas para orientar ações e decisões agrícolas (KAMBLE; GUNASEKARAN; GAWANKAR, 2020).

2.2.2 Big Data

Segundo Waga e Rabah (2014), o *Big Data Analytics* refere-se à capacidade de acumular e analisar grandes volumes de dados estruturados e não estruturados, cujo tamanho excede os limites das ferramentas tradicionais de processamento. Essa tecnologia é capaz de fortalecer a conexão entre agricultores e a indústria de alimentos, promovendo interações duradouras. Os dados estruturados, como tabelas de banco de dados, possuem formato organizado e são facilmente processáveis, sendo amplamente utilizados para decisões empresariais. Já os dados não estruturados, como textos, áudios, imagens e vídeos, embora mais complexos de analisar, oferecem uma fonte rica de informações, especialmente em setores como as redes sociais, onde revelam padrões de comportamento dos consumidores (BRONSON; KNEZEVIC, 2016).

2.2.3 Cloud Computing

A coleta e o processamento diário de grandes volumes de dados exigem ferramentas específicas

de *Big Data* e um alto esforço computacional para garantir eficiência e simplicidade nas operações. Uma etapa essencial nesse processo é o deploy da aplicação, que requer hospedagem em servidores capazes de executar tarefas com alto desempenho e baixo consumo de recursos. No entanto, gerenciar servidores incluindo aspectos como armazenamento, memória, redes e segurança pode ser complexo e oneroso. Nesse cenário, a computação em nuvem (*cloud computing*) surge como solução ideal, permitindo acesso a recursos computacionais sob demanda, com escalabilidade, flexibilidade e sem a necessidade de infraestrutura própria, o que reduz custos, aumenta a eficiência e melhora a segurança dos dados (AMAZON, 2023). De acordo com Bhattarai et al. (2019), a computação em nuvem oferece uma alternativa promissora para lidar com cargas de trabalho intensivas, graças à sua capacidade de operar em paralelo, de forma distribuída, resiliente e segura.

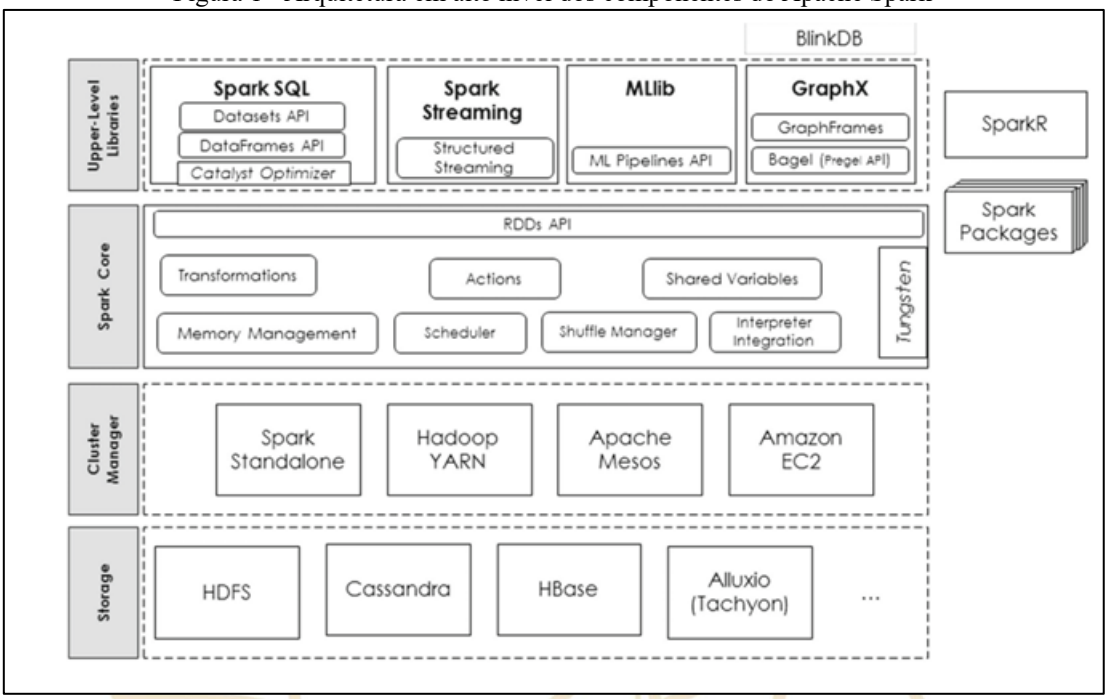
2.2.4 Linguagem Python

O Python tornou-se uma das linguagens mais utilizadas no contexto de *Big Data* devido à sua sintaxe simples, facilidade de uso e amplo ecossistema de bibliotecas especializadas (PYTHON, 2023). Ferramentas como Pandas e NumPy oferecem estruturas de dados eficientes e permitem manipulações complexas de forma intuitiva. Para demandas maiores, bibliotecas como PySpark viabilizam a computação distribuída e o processamento paralelo, acelerando o tratamento de grandes volumes de dados. Além disso, o Python se integra facilmente com tecnologias como Hadoop e Apache Spark, facilitando a construção de pipelines escaláveis. Sua versatilidade também permite aplicações em aprendizado de máquina e inteligência artificial, com o suporte de bibliotecas como TensorFlow, Keras e scikit-learn, tornando-o uma linguagem estratégica para profissionais que atuam em ciência de dados e análise preditiva (PYTHON, 2023).

2.2.5 Processamento paralelo e distribuído com Apache Spark

Com o crescimento exponencial da geração e armazenamento de dados impulsionado pelo *Big Data*, o Apache Spark tem se consolidado como uma ferramenta essencial para o processamento distribuído e escalável de grandes volumes de dados. Diferente do Pandas, que é eficaz para manipulação de dados em um único nó, mas limitado em ambientes distribuídos, o Spark (Figura 1) foi projetado para operar em clusters, executando tarefas paralelas que permitem a escalabilidade horizontal e maior eficiência (ZAHARIA et al., 2016). Uma de suas principais vantagens é o processamento em memória distribuída, que reduz o tempo de resposta e melhora o desempenho com grandes conjuntos de dados.

Figura 1 - Arquitetura em alto nível dos componentes do Apache Spark



Fonte: Salloum et al. (2016).

Além disso, o Spark é compatível com várias linguagens, como Python, Scala, Java e R, e oferece bibliotecas integradas, como Spark SQL, Spark Streaming e Spark MLlib, que ampliam seu uso para análises avançadas, processamento em tempo real e aprendizado de máquina (SALLOUM et al., 2016).

2.2.6 Ferramentas de Armazenamento de dados (AWS S3)

O armazenamento na nuvem é um pilar fundamental da computação em nuvem, servindo como base para aplicações que envolvem *Big Data, Data Warehouses*, Internet das Coisas, bancos de dados e sistemas de backup, oferecendo maior confiabilidade, escalabilidade e segurança em comparação aos sistemas locais tradicionais. Para o projeto em questão, será utilizado o serviço de armazenamento de objetos da Amazon Web Services, o Amazon S3, que é amplamente reconhecido por sua robustez, sendo responsável pelo armazenamento de petabytes de dados com uma impressionante resiliência de 99,99999999999% (AMAZON, 2023). O S3 é amplamente adotado por empresas para a criação de Data Lakes, permitindo armazenar dados semi-estruturados e não estruturados de forma escalável, econômica e segura.

2.2.7 Orquestração de atividades com Apache Airflow

Empresas digitais com aplicações utilizadas por milhões de usuários diariamente geram enormes volumes de dados provenientes de diferentes fontes e funcionalidades. Para organizar e centralizar essas informações em estruturas como Data Lakes ou Data Warehouses, é necessário orquestrar e executar centenas ou até milhares de processos de coleta, transformação e carga de dados. Nesse cenário, o Apache Airflow destaca-se como uma das principais ferramentas open-source para orquestração de fluxos de

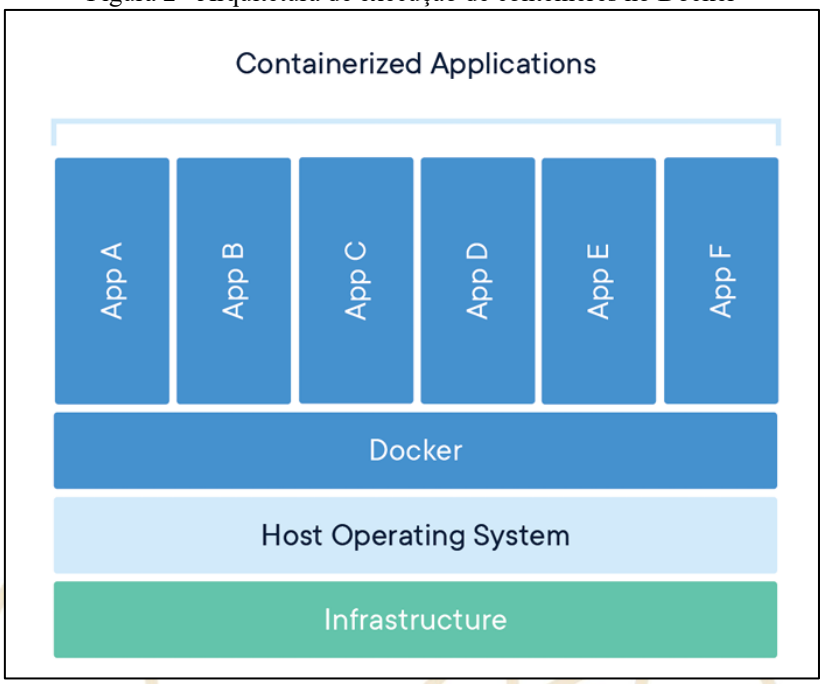
trabalho (Data Pipelines), oferecendo escalabilidade, flexibilidade e uma ampla variedade de operadores para integração com outras tecnologias, como Apache Spark, além de recursos para monitoramento, controle de falhas e agendamento automatizado de tarefas (HARENSLAK; RUITER, 2021).

O uso do Apache Airflow é especialmente valioso em ambientes de *Big Data*, onde há a necessidade de processar dados em diferentes horários, formatos e destinos. Um exemplo prático de sua aplicação seria a construção de um dashboard de dados climáticos, envolvendo a extração de informações de uma API externa, sua transformação e limpeza, e o posterior carregamento em uma interface visual. Com o Airflow, cada uma dessas etapas é configurada como uma tarefa interdependente dentro de um pipeline, permitindo o controle total sobre a execução e a automação do processo de ponta a ponta, com ganhos significativos em eficiência, confiabilidade e escalabilidade (AIRFLOW, 2023).

2.2.8 Conteinerização

Os contêineres representam uma tecnologia de virtualização leve e eficiente, que permite empacotar aplicações junto com todas as suas dependências, bibliotecas e configurações em um ambiente isolado, garantindo que elas sejam executadas de forma consistente em diferentes sistemas e ambientes computacionais (DOCKER, 2023). Essa portabilidade é possível graças às imagens de contêiner, que são criadas a partir de arquivos Dockerfile e podem ser executadas em qualquer host com suporte a contêineres, como o Docker, figura abaixo.

Figura 2 - Arquitetura de execução de contêineres no Docker



Fonte: Docker (2023).

Entre as principais vantagens dos contêineres estão o baixo consumo de recursos, já que compartilham o kernel do sistema operacional do host, e a rapidez no processo de inicialização e parada, o que facilita a escalabilidade e a implantação ágil de aplicações (DOCKER, 2023). Essa abordagem tem se tornado essencial em ambientes modernos de desenvolvimento e operações (DevOps), pois permite criar soluções mais portáteis, estáveis e reproduzíveis, eliminando problemas relacionados a inconsistências de ambiente ou configurações específicas de sistema. Assim, o Docker Daemon é executado acima do Sistema Operacional, e este controla os recursos para os diferentes aplicativos executados em ecossistemas isolados (utilizando conceitos como RunC e ContainerD, comuns para executar virtualização em Linux).

2.2.9 Análise de dados com Power BI

O Power BI, desenvolvido pela Microsoft, é uma das ferramentas mais utilizadas para análise e visualização de dados, oferecendo uma ampla gama de recursos para a criação de dashboards interativos e gráficos variados, como linhas, barras e mapas. Ele permite conexões com diversas fontes de dados, desde planilhas do Excel até bancos de dados e serviços em nuvem, facilitando a criação de filtros e cálculos avançados em um ambiente semelhante ao Excel, o que contribui para sua popularidade entre usuários de diferentes níveis técnicos (MICROSOFT, 2023).

Em ambientes de Big Data, o Power BI se destaca por sua capacidade de se integrar facilmente com plataformas como a AWS, permitindo a visualização e análise de grandes volumes de dados de forma intuitiva e eficiente. Por isso, ele costuma ser utilizado como a etapa final de um pipeline de dados, onde os resultados processados são transformados em insights visuais para a tomada de decisões estratégicas.

3 METODOLOGIA

A utilização de dados tornou-se um recurso cada vez mais estratégico para empresas e organizações de todos os setores. Por meio da análise estruturada de informações, é possível tomar decisões mais embasadas, otimizar processos, antecipar tendências de mercado e identificar oportunidades de melhoria. Nesse cenário, a construção de uma solução orientada a dados exige uma base teórica sólida, fundamentada em pesquisas de artigos científicos, livros, documentações técnicas e trabalhos correlatos.

O primeiro passo para o desenvolvimento do projeto consiste na compreensão clara do problema a ser resolvido com dados, alinhando expectativas, limitações, desafios e objetivos do negócio (conforme ilustrado na parte 1 da Figura 3). Em seguida, é essencial identificar e mapear as principais fontes de dados que serão utilizadas, as quais podem incluir bancos de dados relacionais, arquivos estruturados, relatórios corporativos, dispositivos de campo (IoT) e até mesmo fontes online (parte 3 da Figura 3). Após a identificação, os dados serão armazenados em uma camada bruta, utilizando soluções escaláveis de armazenamento de objetos, como o Amazon S3 ou o Hadoop Distributed File System (HDFS) (parte 2 da Figura 3). O processamento ocorrerá de forma distribuída por meio do framework Apache Spark, aplicando transformações conforme as regras de negócio e estruturando os dados nos formatos Parquet ou Delta na camada analítica (parte 4 da Figura 3).

Por fim, os dados processados serão utilizados para análises avançadas por meio de relatórios e dashboards interativos, empregando ferramentas como Power BI ou Amazon QuickSight. A arquitetura também permitirá o treinamento de modelos de aprendizado de máquina, ampliando o potencial preditivo da solução (parte 5 da Figura 3). Para garantir que a plataforma seja segura, resiliente e eficiente, ela será desenvolvida conforme os princípios do Well-Architected Framework, contemplando aspectos como segurança, desempenho, otimização de custos e sustentabilidade da nuvem (parte 2 da Figura 3):

- Criação da plataforma de nuvem, utilizando Amazon Web Services;
- Adição de usuários de acesso à plataforma de nuvem, respeitando boas práticas como controle granular de acessos, utilizando princípios como o Least Privilege Access (Acesso de menor privilégio);
- Anexação de políticas e funções de controle para garantir acesso apenas aos componentes de nuvem necessários para executar suas funções;
- Criação e controle de VPCs (Virtual Private Cloud), para definição de grupos de segurança, controle de subnets e adição de firewalls;
- Catalogação de dados, para controle de tipagem e volumetria de dados na parte de armazenamento;

- Adição de controle de monitoria de logs das etapas de processamento dos dados, utilizando o CloudWatch.

Figura 3 - Infográfico das 5 principais etapas do projeto

Fonte: Dados da pesquisa, 2023.

Dessa forma, para comprovar os objetivos da pesquisa, este trabalho será fundamentado teoricamente com base em artigos, livros e documentos acadêmicos, com foco em duas áreas principais: Ciência da Computação e Agronomia. Na área de computação, a pesquisa concentra-se em temas como arquitetura de dados, sistemas distribuídos e desenvolvimento de software, com ênfase em ferramentas como Apache Spark e Airflow. Já na Agronomia, o foco está em conceitos como agricultura sustentável, agricultura de precisão, manejo de plantas e produção de alimentos, que representam o núcleo dos dados processados na plataforma. O trabalho abordará o uso de sistemas de *Big Data Analytics* na agricultura moderna, destacando seu potencial para revelar padrões ocultos e promover avanços tecnológicos, sociais e econômicos.

4 RESULTADOS

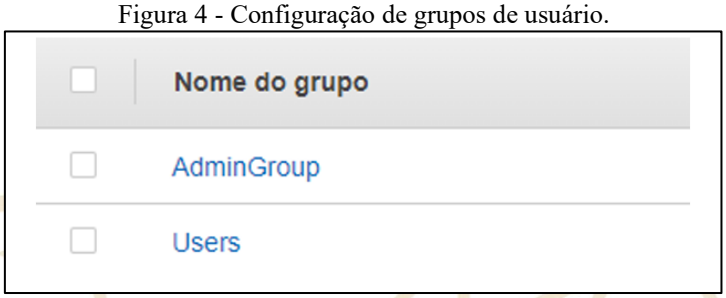
4.1 CRIAÇÃO DE CONFIGURAÇÃO DA CONTA ROOT AWS

A primeira etapa consiste na criação da conta root na AWS, seguindo boas práticas de segurança, como uso de senha forte, ativação de bloqueios para evitar custos excessivos e configuração da autenticação multifator (MFA). Por fim, a AWS recomenda que essa conta não seja utilizada no dia a dia, devido ao seu acesso total aos recursos da plataforma.

4.2 CONFIGURAÇÃO DO USUÁRIO (IAM)

Para garantir o uso seguro e confiável dos serviços da Amazon Web Services, é essencial criar e configurar um usuário de acesso normal com permissões restritas, seguindo o Princípio dos Privilégios Mínimos (Least Privileges Principle). Com base nisso, foi criado o usuário principal de acesso e configurados grupos específicos de permissões para controlar o acesso à plataforma. A Figura 2 ilustra os dois grupos principais criados.

Figura 4 - Configuração de grupos de usuário.



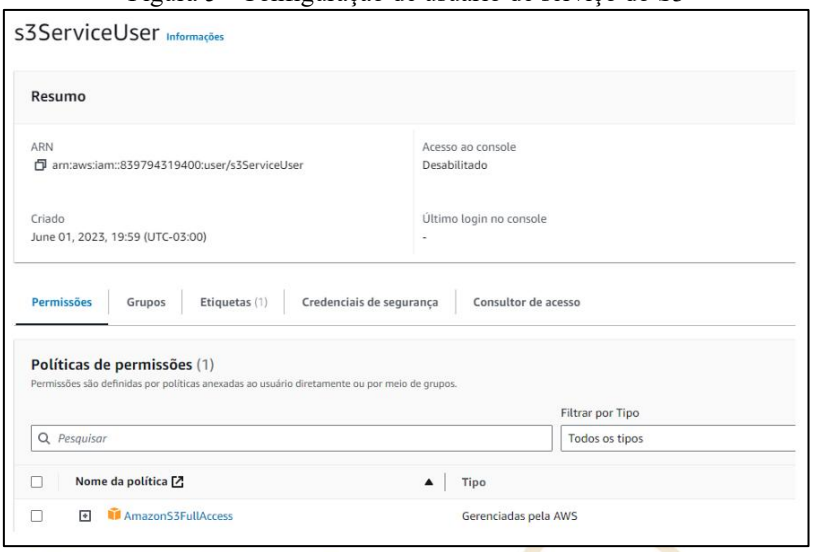
Fonte: Dados da pesquisa, 2023.

O grupo "AdminGroup", possuí a política de permissão de "AdministratorAccess", ou seja, fornece acesso de administrador aos serviços da AWS. Já, o grupo "Users", fornece a política "ReadOnlyAccess", ou seja, apenas acesso à leitura a alguns serviços e processos da AWS. Estas políticas poderão ser mais refinadas para alguns usos mais específicos.

4.3 CRIAÇÃO DE POLÍTICAS E FUNÇÕES PARA SERVIÇOS

Após a configuração de acesso dos usuários à AWS, tornou-se necessário habilitar o acesso programático aos serviços da nuvem, como por meio de APIs ou SDKs. Para isso, foram criados funções e usuários de serviço. Especificamente para o serviço Amazon S3, foi configurado um usuário de serviço com a função "S3FullAccess", conforme ilustrado na Figura 5.

Figura 5 - Configuração de usuário de serviço do S3



Fonte: Dados da pesquisa, 2023.

Esse usuário possui uma chave de acesso e uma senha secreta, que são utilizadas na configuração do Airflow. Com isso, o Airflow pode acessar os scripts de execução e interagir diretamente com os buckets do S3, realizando leitura e escrita de dados de forma automatizada.

4.4 CONFIGURAÇÃO DE VPCS E REDES

Para os serviços se comunicarem seguramente, foi realizada a criação de Virtual Private Cloud (VPCs) e a configuração de alguns aspectos de redes, como o tamanho das subnets (CIDR), tabelas de roteamento, firewall (grupos de segurança), Internet Gateway (para os serviços acessarem a internet), NAT, entre outros serviços.

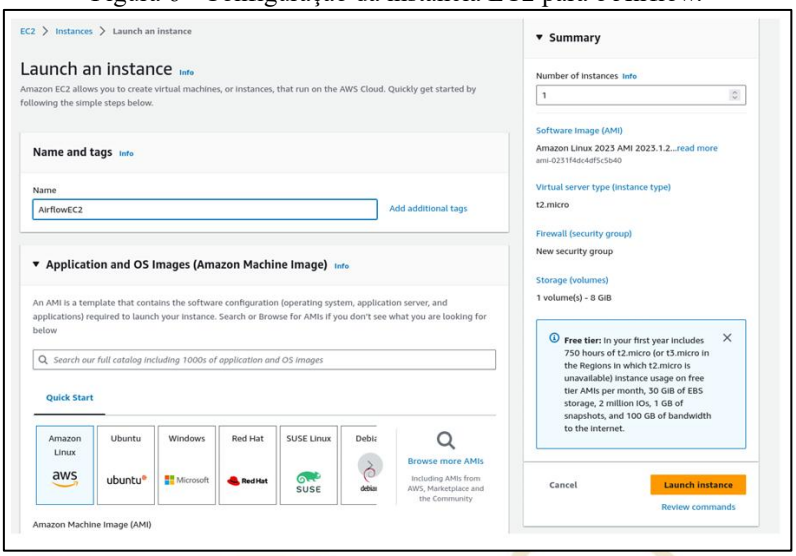
4.5 CRIAÇÃO DO DATA LAKE (BUCKETS NO S3)

Para o armazenamento de dados, foram criados alguns repositórios, chamados Buckets no Amazon Simple Storage Service (S3). Os principais serão os das camadas RAW, CURATED e ANALYTICS. Além disso, foi criado buckets para armazenar scripts e dados de logs. Todos os buckets foram criados na região sa-east-1 (São Paulo).

4.6 CRIAÇÃO DE SERVIDOR EC2 CONTENDO O APACHE AIRFLOW

Para o ambiente de produção, o Airflow será criado em uma instância do Amazon Elastic Compute Cloud (EC2), soluções de máquinas virtuais da AWS. Para criar uma instância do Airflow, será necessário executar os passos conforme a Figura 6. Primeiramente, é necessário criar o cluster EC2 seguindo passo a passo no console da AWS.

Figura 6 - Configuração da instância EC2 para o Airflow.



Fonte: Dados da pesquisa, 2023.

Por fim, é necessário adicionar um código de inicialização, que irá auxiliar a criação e a configuração inicial da máquina virtual para o Airflow.

4.7 INTEGRAÇÃO E DEPLOY DE DAGS E SCRIPTS DO GITHUB NO S3

Além disso, outro ponto importante para a plataforma é a utilização da integração e delivery continuo (em inglês, CI/CD) das Dags e dos códigos Python para dentro do S3. Para isso, será utilizado o GitHub Actions, onde será automatizado o deploy do repositório atualizado em umbucket do S3. Assim, em produção, será sempre obtido os códigos mais atuais.

4.8 CRIAÇÃO E CONFIGURAÇÃO DO CLUSTER ELASTIC MAP REDUCE (EMR)

Para realizar o processamento distribuído dos dados em produção, será utilizado o AWS Elastic Map Reduce. Em produção, este processo será automatizado, e será executado pelo próprio operador do Airflow, o EmrCreateJobFlowOperator. Este e outros operadores criarão o fluxo de criação do cluster com configurações personalizadas, adição das etapas (steps) de processamento, monitoramento de fluxo e logs, e por fim finalização do cluster. Além disso, também há a opção de executar etapas no cluster manualmente utilizando o console da AWS.

4.9 CONFIGURAÇÃO DE CRAWLERS E CATÁLOGO DE DADOS COM AWS GLUE

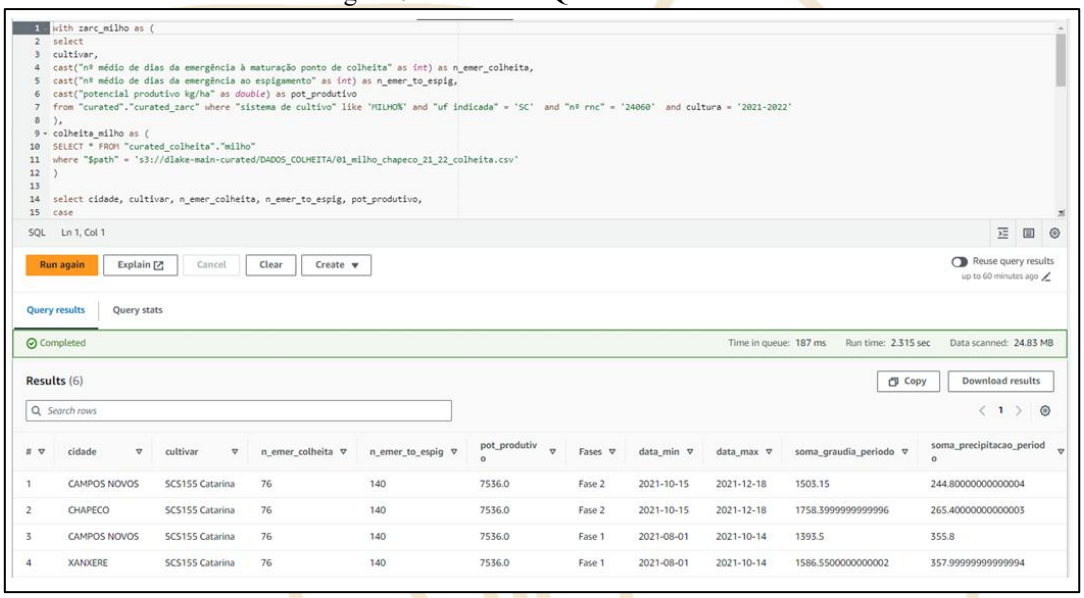
Outra etapa essencial no desenvolvimento de uma plataforma de dados é o controle do catálogo de dados em um data lake, o que permite gerenciar esquemas de tabelas e consultar dados já processados. O AWS Glue oferece essa funcionalidade com uma estrutura semelhante a bancos de dados relacionais, permitindo configurações detalhadas conforme o formato dos arquivos. Ele integra-se com ferramentas como Hive (utilizada no Spark e EMR) e Presto (usada na Athena), possibilitando

consultas eficientes. A inserção de tabelas no Glue pode ser feita manualmente, especificando colunas, tipos e o caminho no S3, ou de forma automatizada por meio dos Crawlers, que identificam e carregam esquemas e partições com base em diretórios definidos e horários programados.

4.10 CONFIGURAÇÃO DO AWS ATHENA PARA REALIZAR QUERIES NO DATA LAKE

Após a realização do catálogo dos dados na camada curated, estes dados estarão automaticamente disponibilizados para buscas e analises no Athena. O AWS Athena é uma Query Engine onde permite o usuário utilizar a linguagem SQL (Structured Query Language) para retornar dados e fazer análises mais profundas (Figura 7).

Figura 7 - Interface SQL do AWS Athena



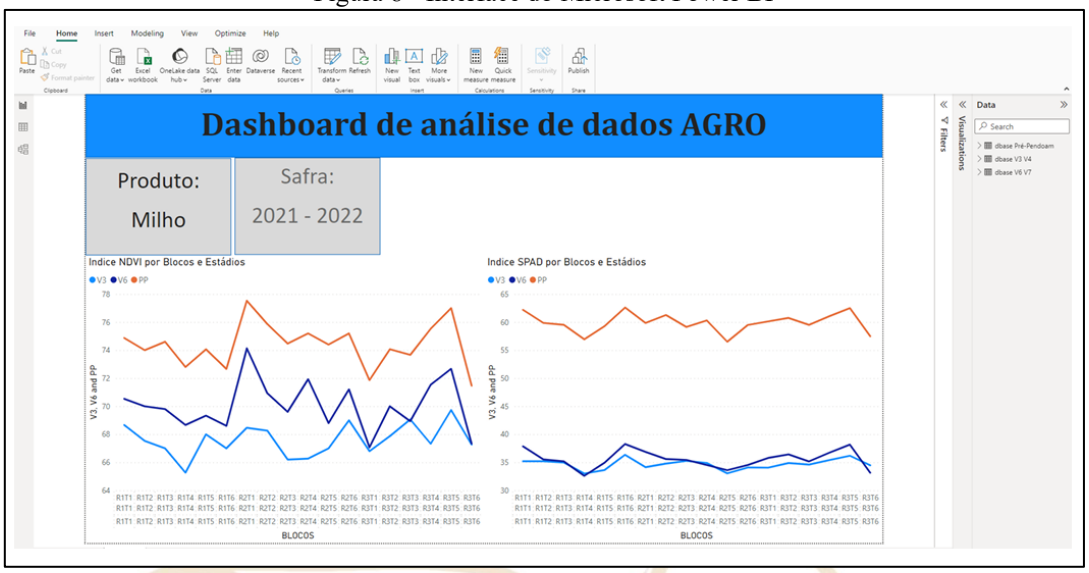
Fonte: Dados da pesquisa, 2023.

Assim, o Athena possui como base a engine e as sintaxes do Presto. A figura acima representa a interface gráfica do Athena, com a execução de queries e a entrega de resultados.

4.11 INTEGRAÇÃO DE DASHBOARDS COM O MICROSOFT POWER BI

Por fim, os dados já curados e catalogados no data lake poderão ser utilizados para análises de negócio mais aprofundadas, com o objetivo de identificar padrões e gerar insights relevantes. Para isso, será utilizada a ferramenta Microsoft Power BI (Figura 8), conhecida por sua capacidade de criar relatórios e dashboards interativos com visualizações intuitivas, cálculos complexos e representações gráficas eficazes.

Figura 8 - Interface do Microsoft Power BI



Fonte: Dados da pesquisa, 2023.

Embora a AWS ofereça uma ferramenta semelhante, o Quicksight, optou-se pelo Power BI neste projeto devido ao seu custo reduzido, já que é gratuito para uso individual, enquanto o Quicksight possui custos mais elevados.

5 DISCUSSÃO

5.1 SOMA TÉRMICA DO MILHO EM GRAUS-DIA

À primeira vista, é fundamental compreender o conceito de soma térmica e sua importância na previsão do crescimento do milho. Assim como outras culturas, como soja e feijão, o milho depende de condições climáticas específicas para se desenvolver adequadamente. Fatores como temperatura, umidade, radiação solar, vento e características do solo influenciam diretamente o crescimento e a produtividade da planta.

Dentre esses fatores, a temperatura exerce papel central. De acordo com Wagner et al. (2013), o milho apresenta alta taxa de crescimento em ambientes com dias quentes e ensolarados, e noites com temperaturas amenas. O intervalo ideal para seu desenvolvimento está entre 24 °C e 30 °C, sendo que temperaturas abaixo de 10 °C ou acima de 30 °C comprometem seu crescimento e, portanto, não devem ser consideradas no cálculo da soma térmica. Em resumo, o milho precisa acumular certa quantidade de energia térmica, expressa em unidades calóricas ao longo de seu ciclo de vida. Esse acúmulo é obtido por meio da soma térmica, que representa a energia necessária para que a planta avance por suas diferentes fases, do plantio à floração.

A fórmula da soma térmica em um dia é correspondida pela equação abaixo:

$$\frac{T_{max} + T_{min}}{2 - T_{base}} \qquad (1)$$

onde:

T_{max}: Temperatura máxima diária; T_{max}: Temperatura mínima diária; T_{max}: Temperatura base (limite inferior) para o crescimento do milho, no caso de 10ºC;

Esta equação retornará um valor correspondente a quantidade de energia térmica obtida durante o dia.

5.2 COLETAS DE FONTES DE DADOS

A principal base de dados utilizada para a análise de informações possui correlação com as análises de campo de algumas colheitas realizadas envolvendo o cultivo de Soja, Trigo, Milho, Feijão, entre outros. Os dados estão normalmente em formato do Excel (.xlsx) e serão processados manualmente e em lotes pela plataforma. Essas informações serão importantes para fazer o cruzamento com as bases de dados complementares, como informações meteorológicas e de solo, entre outras

Além disso, a principal fonte de dados meteorológicos utilizada neste trabalho será o INMET, que disponibiliza uma base abrangente com milhares de estações meteorológicas distribuídas por todo o território brasileiro. Essas estações realizam coletas com frequência horária, fornecendo dados essenciais para a análise climática de regiões específicas. O INMET também oferece, de forma gratuita, a BDMEP (Base de Dados Meteorológicos para Ensino e Pesquisa), que reúne um extenso histórico de dados em alguns casos, com registros de várias décadas permitindo análises de longo prazo. Esses dados são fundamentais para o entendimento do comportamento climático e sua influência no desenvolvimento de culturas agrícolas como o milho.

Dessa forma, o INMET disponibiliza uma base histórica de dados meteorológicos com registros a partir do ano 2000, abrangendo diversas estações localizadas em cidades e estados de todo o Brasil. Esses dados estão acessíveis gratuitamente no portal oficial (https://portal.inmet.gov.br/dadoshistoricos) e são fornecidos em formato CSV, totalizando aproximadamente 9.000 arquivos e 6,5 gigabytes de informações. No projeto, será realizada uma carga batch completa desses dados para a camada raw (dados brutos) e para a camada curated (dados tratados) do data lake da plataforma. Para o processamento e a evolução dos dados, foram utilizados scripts personalizados executados na infraestrutura do Apache Spark, garantindo escalabilidade e eficiência no tratamento dos dados meteorológicos.

Além da base do INMET, será utilizada a API NASA POWER, que fornece dados climatológicos globais por satélite. A coleta ocorrerá de forma horária e diária, utilizando o Apache Airflow e funções Lambda em Python. Embora a base seja completa e com interface intuitiva, não oferece visualização em tempo real nem previsões futuras. Essa fonte complementa

as análises meteorológicas regionais.

Outra fonte essencial para as análises agrícolas é o ZARC (Zoneamento Agrícola de Risco Climático), disponibilizado pelo MAPA. Essa base contém dados sobre cultivares, safras, tempo de crescimento, maturação e potencial produtivo por hectare. Os arquivos, obtidos em formato Excel, são convertidos para CSV para análise na plataforma. Além disso, o ZARC fornece informações sobre riscos climáticos (20%, 30% e 40%) por decênios de plantio, considerando diferentes culturas como milho, soja e trigo. Esses dados são gerenciados pelo MAPA em parceria com a EMBRAPA (EMBRAPA, 2021).

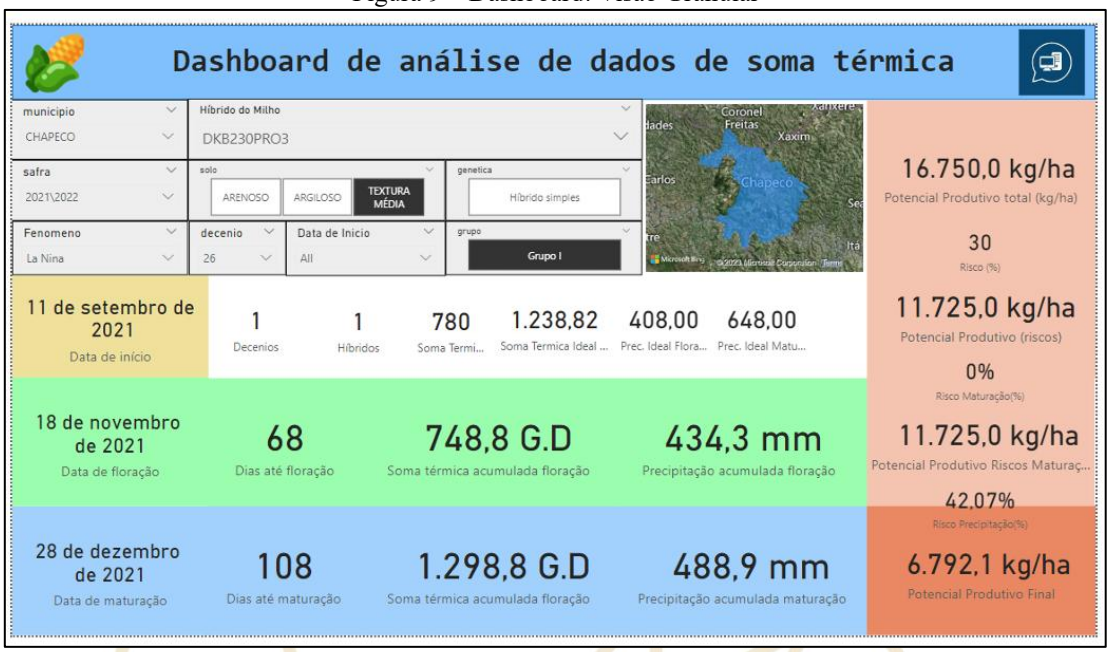
5.3 ANÁLISE DAS INFORMAÇÕES NO POWER BI

Por fim, após a realização da coleta e processamento dos dados de tempo, risco climático e zoneamento agrícola, as informações estarão preparadas para consumo. Os dados na camada analítica foram catalogados utilizando o AWS Glue e distribuídos no AWS Athena. A partir disso, foi utilizado o conector do Microsoft Power BI para AWS Athena, para criação de relatórios e dashboards utilizando as informações analíticas.

Portanto, o Power BI consegue obter os dados inseridos no S3 por meio de queries realizadas no Athena. Isso possibilita a construção de dashboards e relatórios especializados visando entregar as informações de forma simples e intuitiva a usuários. A primeira visão do Dashboard, com informações granulares, pode ser vista na Figura 9.

O detalhamento do dashboard se divide em três partes principais: a parte dos filtros, onde o usuário poderá alterar e ajustar os filtros dos dados; a parte de informações de períodos de produção, a soma térmica acumulada equivalente à cara período e a soma pluviométrica equivalente a cada período, além das informações de soma térmica e pluviométrica ideal, e a parte de análise do potencial produtivo do cultivar considerando os riscos.

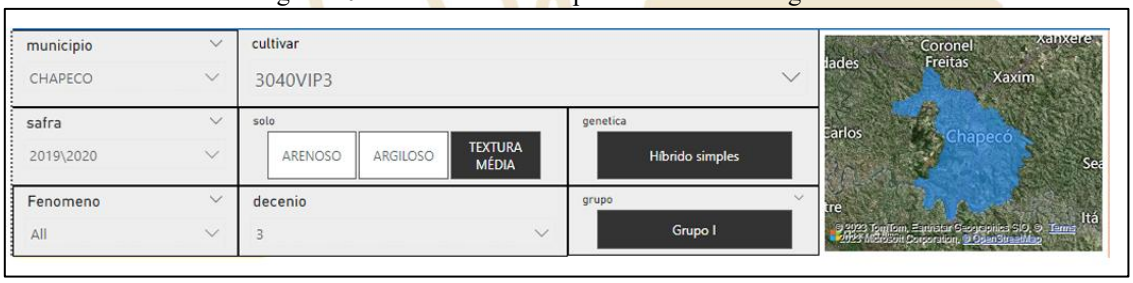
Figura 9 – Dashboard: Visão Granular



Fonte: Dados da pesquisa, 2023.

Esta visão foi criada visando realizar análises granulares com híbridos específicos, em locais e períodos pré-determinados, para obter informações precisas sobre aquele ponto. Há a possibilidade e a disponibilidade do usuário alterar e ajustar configurações e parâmetros para adaptar as condições ideais para realizar análises baseadas nos dados. A parte dos filtros se mantém praticamente inalterados para as outras visões, e pode ser visualizada melhor na Figura 10.

Figura 10 - Dashboard: Principais filtros da visão granular



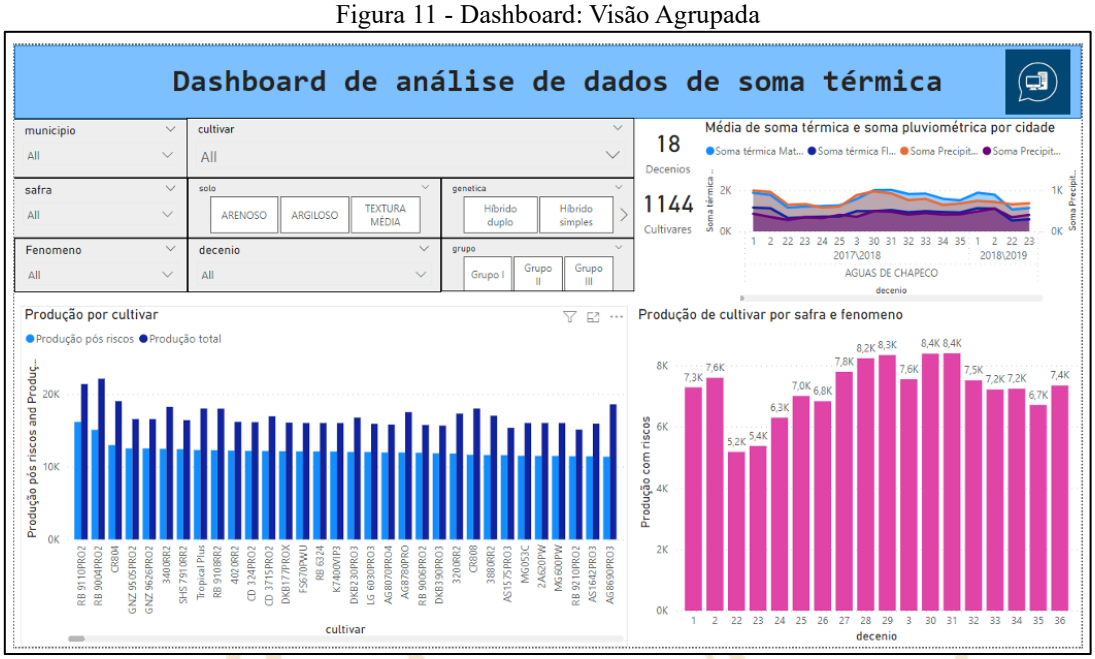
Fonte: Dados da pesquisa, 2023.

O filtro da plataforma é composto por várias opções que permitem ao usuário personalizar a análise dos dados. É possível selecionar o município de plantio (entre 20 disponíveis em SC), a safra desejada ou períodos específicos marcados por fenômenos climáticos como El Niño e La Niña. Também pode-se escolher o decênio de início do plantio, filtros relacionados aos híbridos (como grupo, genética e tipo de solo), além de selecionar um cultivar específico para análise detalhada. Caso algum filtro não seja preenchido, os dados serão apresentados de forma agrupada pela média dos híbridos.

A segunda visão do dashboard (Figura 11) oferece uma análise agrupada dos cultivares,

com gráficos adicionais para facilitar a comparação entre eles. Essa visão permite ao usuário ajustar filtros para identificar, por exemplo, o cultivar com melhor desempenho em uma cidade e safra específicas, considerando riscos climáticos. Também é possível comparar o potencial produtivo de um híbrido em diferentes períodos, utilizando os mesmos filtros da primeira visão, com exceção do filtro granular de cultivar.

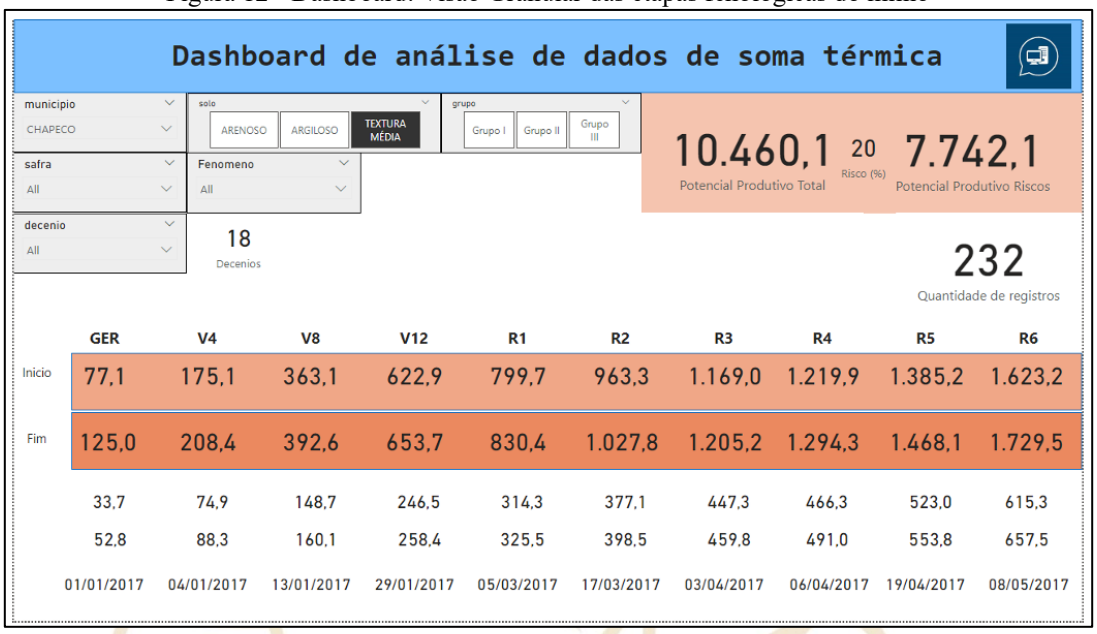
Figura 11 - Dashboard: Visão Agrupada



Fonte: Dados da pesquisa, 2023.

Além das visões anteriores, o dashboard também disponibiliza uma visualização focada nos grupos de tipos de híbridos, com informações granulares por períodos fenológicos do milho (Figura 12). Nessa visão, o usuário pode acompanhar a soma térmica acumulada e a precipitação acumulada em cada fase do crescimento da planta. Os dados são apresentados por meio de KPIs e valores brutos, exigindo uma interpretação mais técnica por parte do usuário para extrair insights relevantes.

Figura 12 - Dashboard: Visão Granular das etapas fenológicas do milho

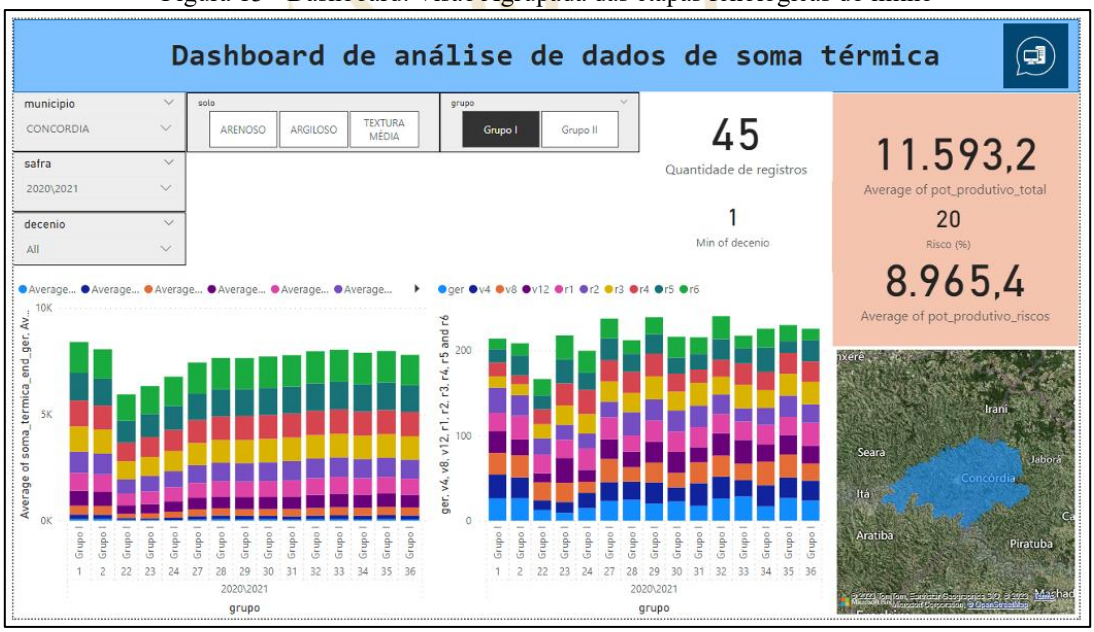


	GER	V4	V8	V12	R1	R2	R3	R4	R5	R6
Inicio	77,1	175,1	363,1	622,9	799,7	963,3	1.169,0	1.219,9	1.385,2	1.623,2
Fim	125,0	208,4	392,6	653,7	830,4	1.027,8	1.205,2	1.294,3	1.468,1	1.729,5
	33,7	74,9	148,7	246,5	314,3	377,1	447,3	466,3	523,0	615,3
	52,8	88,3	160,1	258,4	325,5	398,5	459,8	491,0	553,8	657,5
	01/01/2017	04/01/2017	13/01/2017	29/01/2017	05/03/2017	17/03/2017	03/04/2017	06/04/2017	19/04/2017	08/05/2017

Fonte: Dados da pesquisa, 2023.

Assim como a visão principal, o dashboard também oferece uma visualização com os valores agrupados, apresentada por meio de gráficos de barras (Figura 13). Essa forma mais intuitiva permite ao usuário comparar facilmente as diferenças e quantidades de soma térmica exigidas em cada etapa do desenvolvimento do milho, facilitando a análise e a tomada de decisões agronômicas.

Figura 13 - Dashboard: Visão Agrupada das etapas fenológicas do milho



Fonte: Dados da pesquisa, 2023.

Um aspecto importante da plataforma é a capacidade de realizar análises preditivas baseadas em dados históricos, especialmente considerando os impactos dos fenômenos

meteorológicos La Niña e El Niño, conhecidos por influenciar significativamente a produção dos híbridos. Dessa forma, o usuário pode antecipar quais híbridos terão melhor desempenho na próxima safra (2024-2025, La Niña), com base na média dos resultados observados em anos anteriores sob o mesmo fenômeno. Essa análise futura é apresentada no dashboard por meio das mesmas visões granular e agrupada (Figuras 9 e 11), com filtros e informações pré-configurados para facilitar a consulta.

Portanto, após a construção do dashboard, foram realizadas diversas análises utilizando todos os filtros na classificação granular por híbrido de milho. Para isso, selecionou-se um híbrido específico, em uma localidade determinada (Chapecó), para uma safra e decênio definidos. Também foram considerados o tipo de solo (textura média) conforme o Zarc, além da classificação do híbrido (Grupo 1, Híbrido Simples). Com base nesses parâmetros, a visão final do dashboard apresentou dados de soma térmica e precipitação acumulada para o período analisado. Ao comparar essas informações com análises de campo realizadas por técnicos, bem como dados de outras organizações como EPAGRI e G1, verificou-se que os resultados do dashboard estavam muito próximos às avaliações reais, apresentando uma diferença de apenas 1% no risco climático e 5% no potencial produtivo. Dessa forma, foi possível validar com alta precisão a confiabilidade dos dados e insights fornecidos pela plataforma em relação às condições reais de campo.

6 CONCLUSÕES

A construção e a implementação deste projeto evidenciam sua relevância e potencial na geração de análises e insights de alta qualidade voltados à tomada de decisão no setor agrícola. A plataforma desenvolvida permite a realização de análises com elevado nível de detalhamento, abrindo caminho para aplicações mais avançadas no futuro, como o uso de dados em tempo real e a aplicação de algoritmos de Machine Learning para predições. A capacidade de integrar dados provenientes de múltiplas fontes, incluindo informações meteorológicas e de cultivares em diversos formatos, contribui significativamente para a robustez e a abrangência das análises realizadas.

Apesar dos avanços, algumas etapas previstas inicialmente não foram completamente implementadas. A proposta original incluía a coleta e análise de dados em diversos bancos relacionais em produção; contudo, questões relacionadas à disponibilidade, segurança e uso desses bancos inviabilizaram a migração em tempo real. Além disso, embora o uso da infraestrutura da AWS estivesse previsto para etapas de processamento, essa abordagem foi temporariamente substituída por ambientes de desenvolvimento locais, devido ao alto custo de manutenção de servidores em nuvem. Outro desafio relevante foi a limitação no acesso a bases de dados públicas, já que muitos conjuntos de dados agronômicos possuem restrições de uso ou

caráter confidencial. Ainda assim, o trabalho cumpriu com êxito os objetivos propostos, validando a tese de pesquisa e demonstrando a viabilidade técnica de uma plataforma de dados para a agricultura baseada em *Big Data.*

REFERÊNCIAS

AIRFLOW. Airflow Documentation. 2023. Disponível em: <https://airflow.apache.org/docs/>.

AKHTER, R.; SOFI, S. A. Precision agriculture using iot data analytics and machine learning. Journal of King Saud University-Computer and Information Sciences, Elsevier, v. 34, n. 8, p. 5602–5618, 2022.

AMAZON. AWS Documentation. 2023. Disponível em: <https://docs.aws.amazon.com/>.

BHATTARAI, B. P. et al. Big data analytics in smart grids: State-of-theart, challenges, opportunities, and future directions. IET Smart Grid, Institution of Engineering and Technology, v. 2, p. 141–154, 6 2019. ISSN 25152947.

BRONSON, K.; KNEZEVIC, I. Big data in food and agriculture. Big Data & Society, Sage Publications Sage UK: London, England, v. 3, n. 1, p. 2053951716648174, 2016.

DELGADO, J. A. et al. Big data analysis for sustainable agriculture on a geospatial cloud framework. Frontiers in Sustainable Food Systems, Frontiers Media SA, v. 3, p. 54, 2019.

DOCKER. Docker Documentation. 2023. Disponível em: <https://docs.docker.com/>.

EDWARDS,C. A. Sustainable agricultural systems. [S.l.]: CRC Press, 2020.

EMBRAPA. Milho, Relações com o clima. 2021. <https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/milho/pre-producao/caracteristicas-da-especie-e-relacoes- com-o-ambiente/relacoes-com-o-clima>. Acessado em 14 de novembro de 2023.

HARENSLAK,B. P.; RUITER, J. de. Data Pipelines with Apache Airflow. [S.l.]: Simon and Schuster, 2021.

KAMBLE, S. S.; GUNASEKARAN, A.; GAWANKAR, S. A. Achieving sustainable performance in a data-driven agriculture supply chain: A review for research and applications. International Journal of Production Economics, Elsevier, v. 219, p. 179–194, 1 2020. ISSN0925-5273.

KAMILARIS, A.; KARTAKOULLIS, A.; PRENAFETA-BOLDú, F. X. A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, v. 143, p. 23–37, 2017. ISSN 0168-1699. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169917301230>.

KLEPPMANN, M. Designing Data-Intensive Applications. Beijing: O'Reilly, 2017. ISBN 978-1-4493-7332-0. Disponível em: <https://www.safaribooksonline.com/library/view/ designing-data-intensive-applications/9781491903063/>.105

MICROSOFT. Microsoft Documentation. 2023. Disponível em: <https://learn.microsoft.com/pt-br/docs/>. PYTHON. Python Documentation. 2023. Disponível em: <https://docs.python.org/3/>.

PYTHON. Python Documentation. 2023. Disponível em: https://docs.python.org/3/

SALLOUM,S. et al. Big data analytics on apache spark. International Journal of Data Science and Analytics, Springer, v. 1, p. 145–164, 2016.

WAGA,D.; RABAH, K. Environmental conditions' big data management and cloud computing analytics for sustainable agriculture. World Journal of Computer Application and Technology, Horizon Research Publishing Co., Ltd., v. 2, p. 73–81, 3 2014. ISSN 2331-4982.

WAGNER, M. V. et al. Estimativa da produtividade do milho em função da disponibilidade hídrica em guarapuava, pr, brasil. Revista Brasileira de Engenharia Agrícola e Ambiental, Departamento de Engenharia Agrícola- UFCG, v. 17, n. 2, p. 170–179, Feb 2013. ISSN 1415-4366. Disponível em: <https://doi.org/10.1590/S1415-43662013000200008>.

ZAHARIA, M. et al. Apache spark: a unified engine for big data processing. Communications of the ACM, ACMNewYork, NY, USA, v. 59, n. 11, p. 56–65, 2016.