CARCARÁ: SISTEMA PARA INTEGRAÇÃO DE DADOS METEOROLÓGICOS E HIDROLÓGICOS NO SEMIÁRIDO BRASILEIRO

CARCARÁ: A SYSTEM FOR INTEGRATING METEOROLOGICAL AND HYDROLOGICAL DATA IN THE BRAZILIAN SEMI-ARID REGION

Mikael Mota Feitosa Sales [1]
Orientador: Prof. Dr. José Wellington Franco da Silva [2]
Coorientador: Prof. Dr. Bruno Riccelli dos Santos Silva [3]

RESUMO

O Semiárido Brasileiro demanda acesso integrado a dados meteorológicos e hidrológicos para subsidiar análises ambientais e decisões agrícolas; no entanto, essas informações encontram-se fragmentadas em múltiplas fontes, com heterogeneidade de formatos e baixa padronização. Este trabalho propõe o Carcará, um sistema voltado à integração automática de dados provenientes de fontes como Instituto Nacional de Meteorologia (INMET), Agência Nacional de Águas e Saneamento Básico (ANA) e ERA5-Land (ERA5), atuando na coleta, padronização, armazenamento e disponibilização dessas informações. A solução é estruturada como um pipeline *Extract, Transform, Load (ETL)* composto por ingestão, mensageria, *data lake*, transformação e carga em *data warehouse*, orquestrado por um gerenciador de *workflows*. O sistema prioriza escalabilidade, rastreabilidade e governança, viabilizando a consolidação de grandes volumes de dados com atualização conforme sua disponibilização nas bases de origem. Como resultado, foi gerado um dataset integrado com 17.472 registros horários (192 *timestamps*), cobrindo 44 municípios, 8 UFs e 91 reservatórios, com alinhamento temporal via janela D−6 (defasagem do ERA5); no intervalo, o ERA5 teve cobertura completa, a ANA forneceu nível/volume para 58/91 reservatórios (63,7% dos registros) e o INMET foi o principal limitante (dados em 20/44 municípios e preenchimento meteorológico em 39,6% dos registros), validando a execução sob restrições reais; por fim, as aplicações incluem análises de demanda evaporativa (UR vs. VPD), comparação de UR por UF, fatores da PEV (SSRD, VPD, t2m e WS) e relação PEV vs. umidade observada, entre outras coisas.

Palavras-chave: Arquitetura de dados; Big Data; ETL; Monitoramento Climático; Semiárido Brasileiro.

ABSTRACT

The Brazilian Semi-Arid region requires integrated access to meteorological and hydrological data to support environmental analyses and agricultural decision-making; however, this information is fragmented across multiple sources, with heterogeneous formats and low standardization. This work proposes Carcará, a system aimed at the automatic integration of data from sources such as INMET, ANA, and ERA5, covering data collection, standardization, storage, and availability. The solution is structured as an *Extract, Transform, Load (ETL)* pipeline composed of ingestion, messaging, a *data lake*, transformation, and loading into a *data warehouse*, orchestrated by a *workflow* manager. The system prioritizes scalability, traceability, and governance, enabling the consolidation of large data volumes with updates as they become available in the source repositories. As a result, an integrated dataset with 17,472 hourly records (192 *timestamps*) was generated, covering 44 municipalities, 8 states, and 91 reservoirs, with temporal alignment via a D−6 window (ERA5 lag); within the analyzed period, ERA5 achieved full coverage, ANA provided level/volume data for 58/91 reservoirs (63.7% of records), and INMET was the main

[1] Graduando em Sistemas de Informação na UFC (Campus Crateús) - mikaelmfsales@alu.ufc.br
[2] Professor Doutor em Ciência da Computação na UFC (Campus Crateús) - wellington@crateus.ufc.br
[3] Professor Doutor em Engenharia de Teleinformática na UFC (Campus Crateús) - bruno.silva@crateus.ufc.br

completeness bottleneck (data for 20/44 municipalities and at least one meteorological field filled in 39.6% of records), validating end-to-end execution under real operational constraints; finally, the demonstrated applications include analyses of evaporative demand (RH vs. VPD), comparison of RH by state, drivers of PEV (SSRD, VPD, t2m, and WS), and the relationship between PEV and observed humidity, among others.

Keywords: Data Architecture; Big Data; ETL; Climate Monitoring; Brazilian Semi-Arid Region.

1 INTRODUÇÃO

A tecnologia consolidou-se como elemento estruturante da sociedade contemporânea, influenciando a forma como indivíduos vivem, comunicam-se e aprendem Silva *et al.* (2015). Sua presença é transversal a diferentes setores: na educação, expande o acesso ao conhecimento e potencializa práticas pedagógicas por meio de ferramentas digitais (AKPAN *et al.*, 2025); na saúde, apoia o diagnóstico e a decisão clínica com modelos preditivos (ZHU *et al.*, 2025); e, na indústria, incrementa a eficiência por automação e análise de dados em tempo real (DHANDA *et al.*, 2025).

Apesar dos benefícios associados à adoção de tecnologias em diferentes setores, esse avanço não ocorre de forma uniforme. A exclusão digital ainda impõe barreiras a populações de baixa renda e pequenos empreendedores, limitando o acesso a tecnologias e mantendo práticas manuais pouco eficientes (NEUMEYER *et al.*, 2018; TISIN; OTHMAN, 2024; RAJIANI *et al.*, 2023). Assim, embora a tecnologia seja motor de desenvolvimento, sua incorporação deve ocorrer de forma inclusiva e responsável para mitigar desigualdades e ampliar oportunidades.

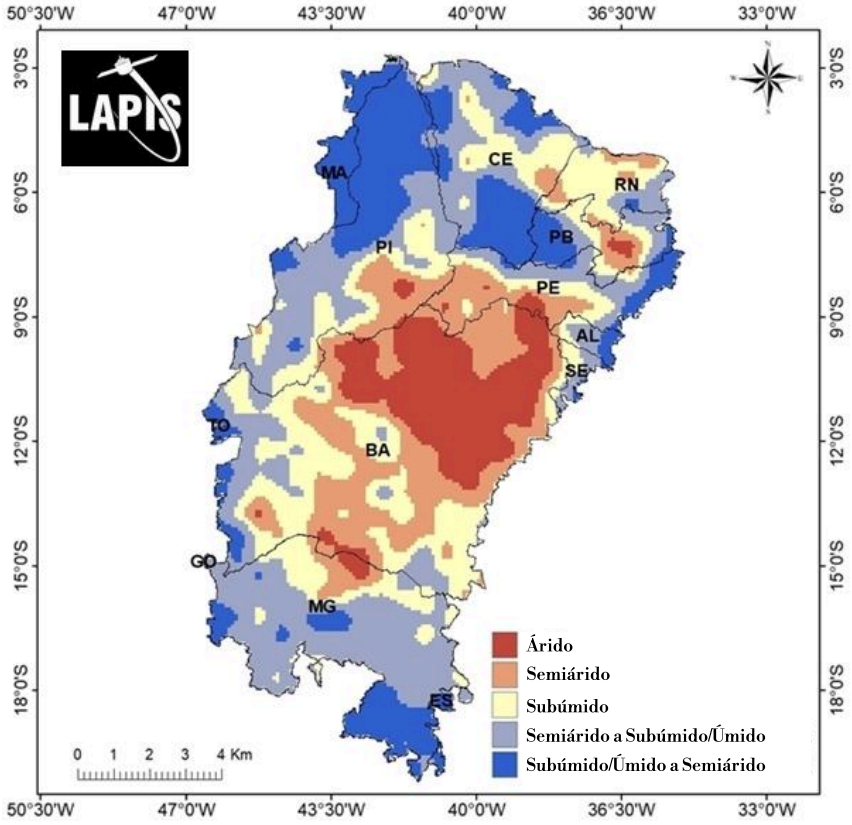
Nesse contexto, a agricultura destaca-se como setor em que a tecnologia exerce papel estratégico, em função de sua relevância socioeconômica e da necessidade de produzir de forma sustentável (MARINELLO *et al.*, 2023; ABIRI *et al.*, 2023). Soluções tecnológicas vêm otimizando etapas do ciclo produtivo, do manejo do solo e definição de períodos de plantio ao monitoramento de pragas, uso racional de insumos e gestão hídrica (MARINELLO *et al.*, 2023; ROPER *et al.*, 2021; JIANG, 2025). Além disso, a adoção de sistemas automatizados, sensoriamento remoto e análise de dados fortalece processos de colheita, logística e comercialização, bem como a capacidade de resposta a eventos climáticos extremos (SHARMA; SHIVANDU, 2024; STEPHEN *et al.*, 2021).

No panorama global, a agricultura é essencial para a segurança alimentar e a estabilidade socioeconômica (VIANA *et al.*, 2021). No Brasil, o setor alimenta aproximadamente 800 milhões de pessoas em escala mundial e apresenta forte participação da agricultura familiar, responsável por cerca de 70% dos alimentos consumidos no país (CONTINI; ARAGÃO, 2021; Câmara dos Deputados, 2023). Sua contribuição econômica também é expressiva: em 2024, respondeu por 23,2% do PIB do Brasil (CEPEA, 2025b) e empregou 28,2 milhões de pessoas (26,02% das ocupações do país) (CEPEA, 2025a). No comércio exterior, de janeiro a novembro de 2024, as exportações do agronegócio somaram US$ 152,63 bilhões, equivalentes a 48,9% do total exportado pelo Brasil no período (PECUÁRIA, 2024).

Entretanto, a eficiência e o desempenho agrícola variam substancialmente conforme condições ambientais e socioeconômicas regionais (PELLEGRINA, 2022) e, em razão de sua diversidade territorial e climática, o Brasil demanda soluções adaptadas às especificidades locais. Nesse cenário, o Semiárido Brasileiro constitui uma região singular: apesar da escassez hídrica e da alta variabilidade climática, abriga cerca de 50% dos estabelecimentos de agricultura familiar do país e exerce papel determinante na segurança alimentar nacional (FRAGOSO *et al.*, 2020). A área semiárida compreende os estados do Nordeste e parte do norte de Minas Gerais, representando cerca de 12% do território nacional e aproximadamente 28 milhões de habitantes (FORTINI; BRAGA, 2020). Seu clima caracteriza-se por precipitação irregular (isoieta de 800 mm anuais), evapotranspiração elevada e índices de aridez que podem ultrapassar 0,50, resultando em risco de seca superior a 60% (INSA, 2023; MARENGO *et al.*, 2020). A Figura 1 apresenta o mapa da classificação climática do Semiárido, evidenciando áreas áridas, semiáridas e subúmidas; embora ciclos de chuvas intensas e concentradas reabasteçam rios e reservatórios,

a dinâmica climática marcada por extremos exige estratégias de gestão adaptativa.

Figura 1 – Mapa da classificação climática do Semiárido.



Fonte: Adaptado de Barbosa (2024)

Diante dessas condições, metodologias orientadas por dados emergem como ferramentas estratégicas para transformar o cenário produtivo. O uso de *big data* na agricultura semiárida permite substituir decisões empíricas por análises quantitativas, integrando múltiplas fontes de informação para orientar o manejo agrícola (MUÑOZ, 2022). Tecnologias de monitoramento ambiental, sensoriamento remoto e sistemas automatizados geram grandes volumes de dados heterogêneos que, quando processados e integrados de forma padronizada, convertem-se em inteligência operacional (AMIRA, 2022).

A adoção de *big data* tem demonstrado benefícios no aprimoramento de modelos de previsão, na otimização de processos agrícolas e no uso racional de recursos naturais (CHEN; LV, 2025). Estudos indicam ganhos na predição de produtividade com o uso combinado de dados climáticos, imagens de satélite e modelos de aprendizado profundo (BHARADIYA *et al.*, 2023; MENA *et al.*, 2024), enquanto a integração de dados geoespaciais, climáticos e de monitoramento em tempo real favorece a otimização de insumos, como água, fertilizantes e energia (WOLANIN *et al.*, 2019). Além disso, sistemas de recomendação e análise preditiva baseados em *big data* fortalecem a tomada de decisão e reduzem a dependência de práticas intuitivas (MUSANASE *et al.*, 2023).

Apesar disso, em regiões de alta vulnerabilidade climática, como o Semiárido, ainda predomina a tomada de decisão agrícola baseada na experiência empírica do produtor. Essa realidade é agravada pela variabilidade interanual das chuvas, longos períodos de seca e elevadas taxas de evapotranspiração, fatores que ampliam incertezas no planejamento e aumentam riscos produtivos (MARENGO *et al.*, 2017). Nesse sentido, a consolidação de uma agricultura orientada

por dados torna-se importante para segurança hídrica, otimização de recursos e maior resiliência produtiva.

Para viabilizar esse avanço, é necessário estabelecer fluxos estruturados de aquisição, pré-processamento e integração de dados, transformando-os em informações úteis para agricultores, gestores e instituições. Assim, sistemas baseados em *big data* configuram um caminho promissor para fortalecer a resiliência e a produtividade agrícola no Semiárido Brasileiro (KAMILARIS *et al.*, 2017; HIMESH *et al.*, 2018; TAURION, 2013).

Nesse contexto, este trabalho propõe o *Carcará*, um sistema para integração automática de dados meteorológicos e hidrológicos, voltado ao recorte climático e agrícola do Semiárido Brasileiro, que enfrenta a fragmentação, a heterogeneidade de formatos e a baixa padronização das bases disponíveis. A proposta organiza um *pipeline* ETL automatizado e de atualização contínua, integrando fontes como INMET, ANA e ERA5-Land, com etapas de coleta, padronização, armazenamento e disponibilização, apoiadas por mensageria, *data lake* e carga em *data warehouse*, priorizando escalabilidade, rastreabilidade e governança.

Este trabalho está estruturado da seguinte forma: a Seção 2 apresenta os conceitos-chave que fundamentam a pesquisa; a Seção 3 discute os trabalhos correlatos; a Seção 4 descreve a concepção e o desenvolvimento do sistema proposto; a Seção 5 detalha o delineamento e a condução dos experimentos; a Seção 6 apresenta e analisa os resultados obtidos com a execução e a aplicação do sistema; por fim, a Seção 7 sintetiza as principais conclusões e indica direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção apresentam-se as bases teóricas necessárias à compreensão do sistema proposto. Inicialmente, discutem-se os conceitos de *Big Data* 2.1 e o processo *Extract, Transform, Load (ETL)* 2.2. Em seguida, abordam-se as arquiteturas de armazenamento *Data Lake* 2.3 e *Data Warehouse* 2.4. Por fim, a Seção 2.5 descreve os bancos de dados e serviços adotados, e a Seção 4.1 apresenta as principais tecnologias utilizadas no desenvolvimento.

2.1 *Big Data*

O conceito de *big data* refere-se a conjuntos de dados cuja escala, complexidade e dinamicidade excedem a capacidade de processamento de abordagens computacionais tradicionais, exigindo modelos distribuídos, arquiteturas escaláveis e técnicas avançadas de análise (KUNE *et al.*, 2016). Na literatura, Li *et al.* (2016), D'Acquisto *et al.* (2015) e Kune *et al.* (2016) caracterizam *big data* partir dos 5Vs:

1. **Volume:** grande escala de dados produzidos continuamente por múltiplas fontes;
2. **Velocidade:** ritmo acelerado de geração, transmissão e processamento, frequentemente em tempo quase real;
3. **Variedade:** heterogeneidade de formatos e estruturas (dados estruturados, semiestruturados e não estruturados);
4. **Veracidade:** confiabilidade, consistência e qualidade, considerando incertezas e ruídos;
5. **Valor:** potencial de extração de *insights* que fundamentem a tomada de decisão.

No contexto agrícola, a produção de dados ocorre de forma massiva e contínua, caracterizando o setor como gerador natural de *big data* (KNEZEVIC, 2016). Essa condição é especialmente explicada por Volume, Variedade e Velocidade: o Volume é impulsionado por sensores ambientais, imagens de satélites, drones, estações meteorológicas e maquinário agrícola, com geração em escalas que podem atingir terabytes por safra (GEWALI *et al.*, 2018);

a Variedade decorre da integração de dados multiespectrais e hiperespectrais, séries temporais, variáveis meteorológicas, parâmetros de solo, informações fenológicas e registros operacionais (KAMILARIS; PRENAFETA-BOLDÚ, 2018); e a Velocidade resulta da transmissão contínua via IoT e redes de comunicação de baixo consumo, como LoRaWAN, NB-IoT e 5G, viabilizando monitoramento e controle em tempo quase real (RAZA *et al.*, 2017).

2.2 *Extract, Transform, Load - ETL*

O processo de *Extract, Transform, Load (ETL)* é um pilar da engenharia de dados, responsável por extrair, transformar e carregar dados em repositórios analíticos, tipicamente um *Data Warehouse (DW)* (KHAN *et al.*, 2024; EL-BASTAWISSY, 2011; MUNAWAR, 2021). Seu objetivo é integrar múltiplas fontes, padronizar e qualificar os dados, garantindo consistência para suporte à decisão.

A extração coleta dados de fontes estruturadas, semiestruturadas e não estruturadas, como bancos relacionais, arquivos (CSV/JSON/XML), APIs, sensores remotos ou *in situ*, imagens, *streams* (ex.: Kafka), planilhas e *web scraping* (QAISER *et al.*, 2023). Nessa etapa, os dados devem ser capturados sem alteração de significado, preservando esquema, tipos e granularidade, com verificações iniciais de disponibilidade e integridade (VASSILIADIS, 2009; KHAN *et al.*, 2024).

A transformação concentra as operações de limpeza, padronização, integração, enriquecimento, normalização e validação, assegurando coerência semântica e qualidade antes da carga no *data warehouse* (KIMBALL; CASERTA, 2004). Inclui remoção de duplicados, tratamento de inconsistências e ausências, padronização de formatos/unidades, integração entre fontes, derivação de atributos, aplicação de regras de negócio e validações temporais/espaciais. Em ambientes analíticos, cerca de 70% do esforço tende a concentrar-se nessa fase (VASSILIADIS *et al.*, 2002).

A carga armazena os dados transformados no repositório analítico conforme o modelo definido (KIMBALL; ROSS, 2013). Pode ser completa (*full load*) ou incremental, preservando histórico e suportando atualizações contínuas. Em arquiteturas modernas, além do *data warehouse* relacional, a carga pode envolver camadas como *Data Marts*, *Lakehouses* e bancos NoSQL para análises específicas.

2.3 *Data Lake*

O *data lake* é uma arquitetura de armazenamento orientada ao depósito centralizado de dados em estado bruto, abrangendo dados estruturados, semiestruturados e não estruturados (SAWADOGO; DARMONT, 2021). Sua característica central é o paradigma *schema-on-read*: os dados são armazenados sem um esquema pré-definido, e a estrutura é aplicada no momento da leitura, oferecendo flexibilidade para diferentes análises e aplicações (DAVENPORT, 2014; MADERA; LAURENT, 2016). A arquitetura de *data lake* é distribuída, fragmentando e distribuindo dados entre múltiplos nós para garantir desempenho e resiliência (SARANYA *et al.*, 2025).

O *data lake* utiliza *Object Storage* em vez de sistemas de arquivos POSIX. A leitura e escrita ocorrem via APIs (`PUT`, `GET`, `DELETE`) e os dados são organizados em *buckets* e *prefixes*, facilitando integração com ferramentas de análise distribuída (RAVAT; ZHAO, 2019). Para suportar governança, descoberta e padronização de acesso, emprega-se um catálogo de metadados (p.ex., *AWS Glue*, *Apache Hive Metastore*, *Unity Catalog*), que descreve esquemas, partições e localizações físicas, viabilizando consultas por motores SQL distribuídos (p.ex., Apache Spark, Trino, Presto) (SAWADOGO; DARMONT, 2020; MEGDICHE *et al.*, 2021;

CHERRADI *et al.*, 2023).

2.4 *Data Warehouse*

O *data warehouse* é um repositório centralizado e consistente (IBM, 2025), definido como uma "coleção de dados orientada por assunto, integrada, variante ao longo do tempo e não volátil, usada primariamente na tomada de decisões organizacionais" (INMON, 2005; CHANDRA; GUPTA, 2018). Sua finalidade é sustentar decisões estratégicas ao consolidar dados de múltiplas fontes e preservar histórico (MARCH; HEVNER, 2007).

A arquitetura do *data warehouse* é frequentemente descrita em três camadas. A Camada de Armazenamento corresponde ao banco do *data warehouse* (PALOPOLI *et al.*, 2002); nela, dados de múltiplas fontes passam por ETL ou ELT para limpeza, transformação e organização antes da incorporação ao armazém (IBM, 2025). A Camada de Processamento Analítico (servidor OLAP) viabiliza agregações e consultas multidimensionais, abstraindo a complexidade do repositório (HAN, 1997). Por fim, a Camada de Apresentação oferece interfaces para relatórios, *dashboards* e análises customizadas (QIANG; LIU, 2009).

Quanto às distinções, *data warehouse* e *data lake* têm propósitos e arquiteturas diferentes (NAMBIAR; MUNDRA, 2022). O *data warehouse* armazena dados estruturados e altamente processados sob *schema-on-write*, com esquema aplicado antes da carga (tipicamente via ETL) (NEHA, 2024; FIDALGO, 2024; WREMBEL, 2011). O *data lake* armazena dados predominantemente brutos sob *schema-on-read*, aplicando estrutura na consulta (HAI *et al.*, 2021). Enquanto o *data warehouse* é orientado a objetivos analíticos específicos e governança mais rígida, o *data lake* favorece ingestão ágil e experimentação *ad hoc* (NAMBIAR; MUNDRA, 2022). Em termos de custo, o *data warehouse* tende a ser mais oneroso pelo nível de estrutura e desempenho, ao passo que o *data lake* prioriza armazenamento de menor custo, como Hadoop e nuvem (RELLA, 2025). Em síntese, o *data warehouse* assemelha-se a uma garrafa de água pronta para consumo, enquanto o *data lake* corresponde a um lago de dados em estado mais natural (DIXON, 2010).

2.5 Bancos de dados

Nesta parte, são descritas as principais fontes de dados que fundamentam o sistema proposto, destacando seu papel na aquisição, cobertura espacial e temporal, e formas de disponibilização. Para isso, apresentam-se: (i) o INMET, como provedor oficial de observações meteorológicas de superfície no Brasil; (ii) a ANA e o Sistema de Acompanhamento de Reservatórios (SAR), como base para o monitoramento hidrológico e o acompanhamento operacional de reservatórios, especialmente no Nordeste e Semiárido; e (iii) o *Copernicus Climate Change Service (C3S)* e o ERA5-Land, como reanálise climática de alta resolução para variáveis atmosféricas, energéticas, do solo e hidrológicas. Em conjunto, essas fontes fornecem subsídios para compor séries históricas das variáveis de interesse, no recorte geográfico e temporal pertinente a esse trabalho.

2.5.1 Instituto Nacional de Meteorologia - INMET

O INMET é o serviço meteorológico oficial do Brasil, vinculado ao MAPA, responsável por monitorar e prever tempo e clima e emitir alertas meteorológicos (INMET, 2025). Atua em todo o território nacional, representa o país na Organização Meteorológica Mundial - OMM e integra a vigilância meteorológica sul-americana. Opera uma rede com cerca de 700

estações de superfície (aproximadamente 560 automáticas e 133 convencionais) com observações horárias e transmissão em tempo quase real de estações automáticas e com medições três vezes ao dia das estações convecionais. A coleta de dados é feita majoritariamente por sensores *in situ*, complementada por satélites e radares para monitoramento e previsão de curto prazo (INMET, 2025). Os dados são disponibilizados via Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) [4] para séries históricas e via Portal do INMET [5] para acesso em tempo real e produtos operacionais (INMET, 2025).

2.5.2 Agência Nacional de Águas e Saneamento Básico (ANA) e Sistema de Acompanhamento de Reservatórios (SAR)

A ANA é uma autarquia federal vinculada ao Ministério da Integração e do Desenvolvimento Regional, responsável por políticas de segurança hídrica, saneamento básico e gestão integrada de bacias hidrográficas. Entre suas atribuições estão o monitoramento hidrológico nacional, a coordenação da Rede Hidrometeorológica Nacional, a outorga e fiscalização do uso da água, a regulação de barragens e a definição de normas de referência para o saneamento (ANA, 2025).

Para apoiar essas atividades, a agência mantém o SAR, desenvolvido a partir de 2013 e lançado em 2014, que centraliza dados operacionais de reservatórios (volume, vazões, capacidade útil, cota e alertas) e se organiza nos módulos SIN, Nordeste e Semiárido e Outros Sistemas Hídricos (ANA, 2025). O módulo Nordeste e Semiárido reúne dados de mais de 500 reservatórios nos nove estados do Nordeste e no norte de Minas Gerais, com informações como estado, bacia, município, capacidade, cota, volume, percentual e data de referência; os dados são públicos no portal [6], com visualização em mapas e séries temporais e *download* em formatos abertos (CSV, JSON e KMZ), permitindo acompanhamento em tempo quase real e reforçando a gestão participativa, especialmente em áreas sujeitas à seca (ANA, 2025).

2.5.3 Copernicus Climate Change Service (C3S) e ERA5-Land

O *Copernicus Climate Change Service (C3S)* é um serviço operacional da União Europeia, coordenado pela Comissão Europeia e implementado pelo *European Centre for Medium-Range Weather Forecasts (ECMWF)*. O C3S integra observações de satélites, medições *in situ* e modelos numéricos por meio de reanálises climáticas, gerando séries temporais homogêneas com alta resolução espacial e temporal. Os produtos são disponibilizados no *Climate Data Store (CDS)*, com acesso gratuito via interface web e API.

O *ERA5* é um produto de reanálise de superfície terrestre do ECMWF no âmbito do Copernicus Climate Change Service (C3S). Ele refina o ERA5 global com foco em processos de superfície e solo, oferecendo resolução espacial de aproximadamente 9 km ($0.1^\circ \times 0.1^\circ$) e resolução temporal horária, cobrindo de 1950 até o presente (MUñOZ-SABATER *et al.*, 2021). A atualização ocorre com atraso operacional típico de poucos dias, mantendo utilidade operacional e qualidade científica.

O ERA5 disponibiliza variáveis geofísicas para modelagem terrestre, hidrológica e climática, incluindo:

- **Atmosfera em superfície**: temperatura do ar a 2 m, pressão à superfície, vento a 10 m, ponto de orvalho e precipitação total;

[4] https://bdmep.inmet.gov.br
[5] https://portal.inmet.gov.br
[6] https://www.ana.gov.br/sar

- **Balanço de energia**: radiação de onda curta e longa (descendente e ascendente), fluxos de calor sensível e latente, albedo e radiação líquida;
- **Solo**: temperatura e umidade do solo em quatro camadas (0–7 cm, 7–28 cm, 28–100 cm e 100–289 cm), água/neve e variáveis associadas;
- **Hidrologia superficial**: escoamento, infiltração, evapotranspiração, drenagem subterrânea e saturação do solo.

Os dados são acessíveis gratuitamente via CDS [7], com consultas e *download* pela web ou acesso automatizado via `cdsapi`. Os arquivos são disponibilizados em NetCDF4 e GRIB, seguindo padrões do *Climate and Forecast (CF) Metadata Convention* para interoperabilidade.

[7] https://cds.climate.copernicus.eu

3 TRABALHOS RELACIONADOS

Stephen *et al.* (2021) discute *datasets* abertos recentes derivados de satélites Landsat, MODIS e *Sentinel*, que reúnem informações de solo, clima, água e vegetação relevantes para problemas agrícolas. O trabalho propõe uma arquitetura de Big Data (BD) baseada em nuvem para armazenamento e processamento desses conjuntos de dados diretamente na infraestrutura do provedor, reduzindo a necessidade de transferência para ambientes locais. Nessa arquitetura, o processamento é elástico, escalando conforme a demanda, e visa apoiar a geração de *insights* para agricultores ao longo das etapas da produção. Para tornar os dados úteis, o estudo descreve um fluxo que transforma produtos brutos (nível 1 – L1) em produtos derivados (nível 2 – L2). Esse fluxo inclui pré-processamento das imagens de grande extensão por particionamento em unidades menores (telhas) e padronização de resolução, viabilizando análises mais específicas por área. Em seguida, realiza-se a integração de séries temporais e de múltiplas fontes, combinando observações passadas e atuais para compor bases consistentes. A extração de informação concentra-se em atributos-chave, como uso e cobertura do solo, condições hídricas e meteorológicas e índices de vegetação. Como resultado, o trabalho apresenta a arquitetura de um sistema capaz de construir perfis temporais de áreas agrícolas a partir de sensoriamento remoto, permitindo caracterizar condições ambientais e de cultivo ao longo do tempo. O sistema prevê o uso de mineração de dados espaço-temporais e modelos de aprendizado profundo para identificar padrões e produzir indicadores aplicáveis. Também é descrita uma interface na qual o usuário define áreas de interesse e obtém saídas analíticas para apoiar decisões como seleção de culturas, manejo de água e aumento de produtividade. Por fim, os autores explicitam que a contribuição é a definição arquitetural, sem implementação prática do sistema Stephen *et al.* (2021).

Wei *et al.* (2023) propõe um sistema de monitoramento da segurança do ambiente de produção agrícola para superar limitações de métodos tradicionais, como alto custo, manutenção complexa e baixa escalabilidade. A solução integra Inteligência Artificial (IA), computação em nuvem e redes de BD para facilitar coleta e transmissão de dados entre equipamentos agrícolas, visando otimizar o uso de recursos, elevar produtividade e qualidade e reduzir impactos ambientais associados ao uso excessivo de insumos. A arquitetura é organizada em três camadas: (i) aplicação, responsável pela interação do usuário com os dados; (ii) sensoriamento, composta por nós distribuídos para coleta em tempo real; e (iii) monitoramento em nuvem, que centraliza armazenamento e processamento. A camada de sensoriamento utiliza WSNs para monitorar variáveis como temperatura, umidade e condições das culturas, com comunicação por *Wi-Fi*, ZigBee e *Bluetooth*. Para localização dos sensores, o trabalho adota o algoritmo *DV Hop*, aprimorado com RSSI para aumentar a precisão de posicionamento. A análise ocorre majoritariamente na nuvem, que processa grandes volumes de dados de sensores meteorológicos, de solo e de cultivo. Algoritmos de IA são aplicados para detecção de padrões, correlações e anomalias, apoiando a tomada de decisão. O estudo também propõe um método para estimar atraso de serviço considerando coordenadas dos nós, tempo de processamento e tempo de espera, com foco em reduzir latência. Como resultados reportados, a integração DV Hop + RSSI melhora a acurácia de localização (90,5% com 120 nós, contra 57,3% no método tradicional), enquanto o uso de nuvem fornece elasticidade e pagamento sob demanda. O consumo energético é tratado por um desenho que favorece transmissão indireta e operação em modo de suspensão quando inativo. Em síntese, o sistema combina WSNs e processamento em nuvem para monitoramento e resposta mais rápida a eventos, com potencial de apoiar gestão agrícola mais eficiente e ambientalmente responsável Wei *et al.* (2023).

O estudo de Garg e Alam (2023) apresenta um sistema de coleta de dados agrícolas baseado em Internet of Things (IoT), capaz de capturar informações em tempo real no campo por

meio de um mini robô (*microbit*) equipado com sensores. Os dados coletados são transmitidos via *Wi-Fi* e armazenados na nuvem, utilizando a plataforma *ThingSpeak*[8], para posterior processamento e análise. A arquitetura é organizada em quatro camadas: *sensing*, *gateway*, *cloud* e *application*. A camada de sensoriamento realiza a aquisição em campo de variáveis como temperatura, umidade do solo e do ar e indicadores de crescimento bacteriano/microbiano. A camada *gateway* provê conectividade e encaminha os dados para a camada em nuvem, na qual o *ThingSpeak* é utilizado para criar canais de armazenamento independentes por parâmetro. A camada de aplicação disponibiliza as leituras ao usuário final. O sistema realiza verificação por limiares para inferir condições de saúde da cultura, e o processamento/visualização é conduzido via recursos de análise do *MATLAB* integrados ao *ThingSpeak*. Os autores mencionam a divulgação de informações e recomendações ao agricultor por meio do *Twitter*, com suporte a análises descritivas, preditivas e prescritivas. Como limitações, o estudo aponta dependência de infraestrutura de Internet (qualidade e custo), além de barreiras de adoção associadas à familiaridade dos usuários com mídias sociais. Para mitigá-las, sugere treinamento e conscientização, *workshops*, acesso a clientes mais baratos (p.ex., *smartphones*), conectividade mais ampla e interfaces de aplicativos orientadas à usabilidade no contexto rural Garg e Alam (2023).

O artigo Rahman *et al.* (2023a) propõe uma estrutura que integra IoT, mineração de dados e monitoramento em nuvem para apoiar agricultura de precisão, com foco em aumento de produtividade, gestão hídrica e previsões de colheita em tempo real via aplicações *web* e *mobile*. A solução utiliza ESP8266/NodeMCU com sensores como DHT11 e sensores de umidade do solo, enviando leituras para a nuvem para armazenamento, processamento e análise. A arquitetura é organizada em sete camadas: *physical* (sensores e atuadores), *link* (conectividade *Wi-Fi*), *encapsulation* (integração do NodeMCU com IPv6 e mecanismos de segurança), *middleware* (coleta e encaminhamento), *configuration* (processamento e roteamento dos dados), *management* (integração, análise e geração de previsões/relatórios) e *application* (apresentação ao usuário). Além do monitoramento, o sistema controla a irrigação por válvulas solenoides, decidindo a ativação/desativação com base em dados e modelos de aprendizagem automática, gerando *insights* sobre saúde das culturas, anomalias e estratégias de manejo. Os dados passam por tratamento com normalização e agrupamento via *K-Means*. O *dataset* combina coleta em tempo real com fontes externas (p.ex., anuário agrícola de Bangladesh 2020 e artigos sobre colheitas), seguido de curadoria e modelagem para extração de padrões. Para validação, os autores aplicam classificadores (DT, NB, MLP e KNN) com validação cruzada *K-fold*, reportando acurácia de 87,3786% na análise Rahman *et al.* (2023a).

Os trabalhos analisados abordam a coleta, o tratamento e a disponibilização de dados para suporte à decisão no contexto agrícola e ambiental, indicando que sistemas embarcados, *frameworks* e técnicas analíticas viabilizam a extração de *insights*, mas também revelando lacunas recorrentes à luz dos critérios adotados neste estudo. Observa-se, em Stephen *et al.* (2021), predominância de sensoriamento remoto e uma contribuição essencialmente arquitetural, o que reforça a necessidade de integrar e validar tais informações com dados *in situ* (IoT/sensores em campo) para ampliar precisão e confiabilidade e subsidiar modelos de aprendizado de máquina; em Garg e Alam (2023), a cobertura restrita de variáveis (temperatura, umidade do ar e do solo e indicadores bacterianos) sugere a ampliação da instrumentação (pH, luminosidade e nutrientes) e a superação de limitações de disponibilização associadas ao ecossistema da plataforma; e, em Rahman *et al.* (2023b), evidenciam-se restrições de disponibilidade/qualidade de dados produtivos e limitações operacionais decorrentes da dependência de *Wi-Fi*, indicando a adoção de alternativas de conectividade (GPRS), maior robustez física dos dispositivos e incorporação sistemática de dados de rendimento. De modo geral, as abordagens investigadas

[8] https://thingspeak.com/

contemplam apenas subconjuntos dos requisitos de integração de fontes e camadas do fluxo de dados (RAHMAN *et al.*, 2023a), ao passo que a proposta Carcará distingue-se por atender, de forma concomitante, aos critérios estabelecidos, integrando dados meteorológicos, hidrológicos e de reanálises/satélites em um *pipeline* automatizado e de atualização contínua, com pré-processamento e integração espaço-temporal e arquitetura baseada em *data lake* e *data warehouse*, orientada ao recorte climático e agrícola do Semiárido Brasileiro.

Este estudo visa abordagens inovadoras e soluções eficientes e se propõe preencher os pontos não cobertos pelos anteriores aqui abordados. A Tabela 1 apresenta os pontos observados comparando cada trabalho ponto a ponto:

Pontos observados:

1. Coleta de dados *in situ* via sensores em campo;
2. Aquisição de dados de sensoriamento remoto (satélites);
3. Coleta de dados hidrológicos (reservatórios, vazão e nível das bacias);
4. Automação do pipeline de ingestão e atualização contínua dos dados;
5. Pré-processamento: limpeza, padronização, enriquecimento e integração espaço-temporal;
6. Arquitetura e governança de dados baseada em Data Lake e Data Warehouse;
7. Disponibilização pública dos dados (repositórios, APIs, etc.);
8. Aderência ao contexto climático e agrícola do Semiárido Brasileiro.

Tabela 1 – Comparação das contribuições entre os trabalhos relacionados e a proposta Carcará

Autor	**Contribuições**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Stephen *et al.* (2021)		X		X	X	X		
Wei *et al.* (2023)	X			X	X		X	
Garg e Alam (2023)	X						X	
Rahman *et al.* (2023a)	X			X	X			
Este trabalho	X	X	X	X	X	X	X	X

Fonte: Elaborado pelo autor.

4 METODOLOGIA

Esta sessão descreve o procedimento metodológico empregado na implementação e operacionalização do sistema proposto, denominado Carcará. Apresentam-se, em sequência, as tecnologias usadas na implementação, arquitetura adotada, fluxo de execução do pipeline *Extract–Transform–Load* e a estratégia de orquestração utilizada para automatizar e coordenar as etapas de coleta, transformação e carga dos dados.

4.1 Tecnologias ultilizadas

Esta subseção descreve as principais tecnologias empregadas na implementação do sistema, explicitando suas funções na arquitetura e no *pipeline* de dados desde a ingestão e integração entre componentes até o armazenamento final.

4.1.1 MinION

O MinIO foi utilizado como camada de *object storage* para compor a infraestrutura de *data lake*, em virtude de sua compatibilidade, desempenho e escalabilidade horizontal (SAVEINCLOUD, 2023; IBM, 2023). No sistema, o MinIO armazenou dados provenientes de múltiplas fontes em seu formato original, com organização em zonas lógicas (*raw*, *refined* e *curated*). O armazenamento foi estruturado em *buckets* e *prefixes*, com acesso por operações REST (`PUT`, `GET`, `DELETE`) e metadados associados aos objetos (MINIO, 2025; REPOSITORYMINIO, 2025).

4.1.2 PostgreSQL

O PostgreSQL foi empregado como SGBD objeto-relacional para a camada analítica, devido à confiabilidade, integridade e aderência a padrões, com suporte a transações ACID e extensibilidade (PostgreSQL Global Development Group, 2024). No sistema, o PostgreSQL recebeu dados pré-processados e harmonizados provenientes da camada de transformação, permitindo armazenamento estruturado e consultas analíticas sobre os dados integrados.

A camada de *data warehouse* foi estruturada com modelagem dimensional em esquema estrela, por sua adequação a ambientes analíticos e suporte a consultas multidimensionais (WARNARS, 2010; WIJAYA *et al.*, 2024). Nesse modelo, a organização foi centrada em uma tabela de fatos conectada diretamente a tabelas de dimensão, favorecendo clareza semântica, redução de junções em consultas recorrentes e usabilidade para análise (WARNARS, 2010; WIJAYA *et al.*, 2024). O esquema estrela foi caracterizado pela presença de uma tabela de fatos contendo métricas e chaves estrangeiras, e tabelas de dimensão responsáveis por contextualizar os fatos com atributos descritivos e hierárquicos Wijaya *et al.* (2024). Essa estrutura foi empregada para otimizar o desempenho de consultas em cenários OLAP, reduzindo complexidade relacional e ampliando a interpretabilidade dos resultados (WARNARS, 2010).

4.1.3 Mensageria – Kafka

A integração entre componentes foi suportada por um serviço de mensageria, viabilizando comunicação assíncrona e desacoplada entre produtores e consumidores, por meio de abstrações de filas e tópicos. Esse modelo permitiu transporte de eventos com persistência, tolerância a falhas e semânticas configuráveis de entrega. Para essa finalidade, foi utilizado o Apache Kafka, plataforma distribuída de mensageria e *streaming* projetada para alta taxa de

transferência, particionamento, replicação e durabilidade Apache Software Foundation (2024). No Kafka, os dados foram organizados em tópicos particionados, armazenados como logs. *Producers* publicaram eventos nos tópicos, enquanto *consumers* realizaram leitura e processamento, podendo organizar-se em grupos. Esse arranjo sustentou a construção de pipelines orientados a eventos, mantendo ordenação por partição e durabilidade das mensagens.

4.1.4 Orquestração de Workflows e Apache Airflow

A coordenação do *pipeline* foi realizada por um mecanismo de orquestração de *workflows*, que permitiu modelar processos complexos como DAGs e controlar dependências e execução de tarefas de forma centralizada. Para isso, foi utilizado o *Apache Airflow*, que operacionalizou o paradigma de *workflow as code*, viabilizando agendamento, monitoramento e rastreabilidade de execuções.

A arquitetura do *Airflow* foi composta por *Scheduler*, *DAG Processor*, *Webserver*, repositório de DAGs e banco de metadados, registrando histórico e estado das tarefas. Em cenários de maior demanda, o processamento foi suportado por *workers* e pelo *executor* configurado, mantendo observabilidade por meio de logs e controle de execução. Dessa forma, o *Airflow* coordenou as etapas de ingestão, transformação e carga.

4.2 Arquitetura

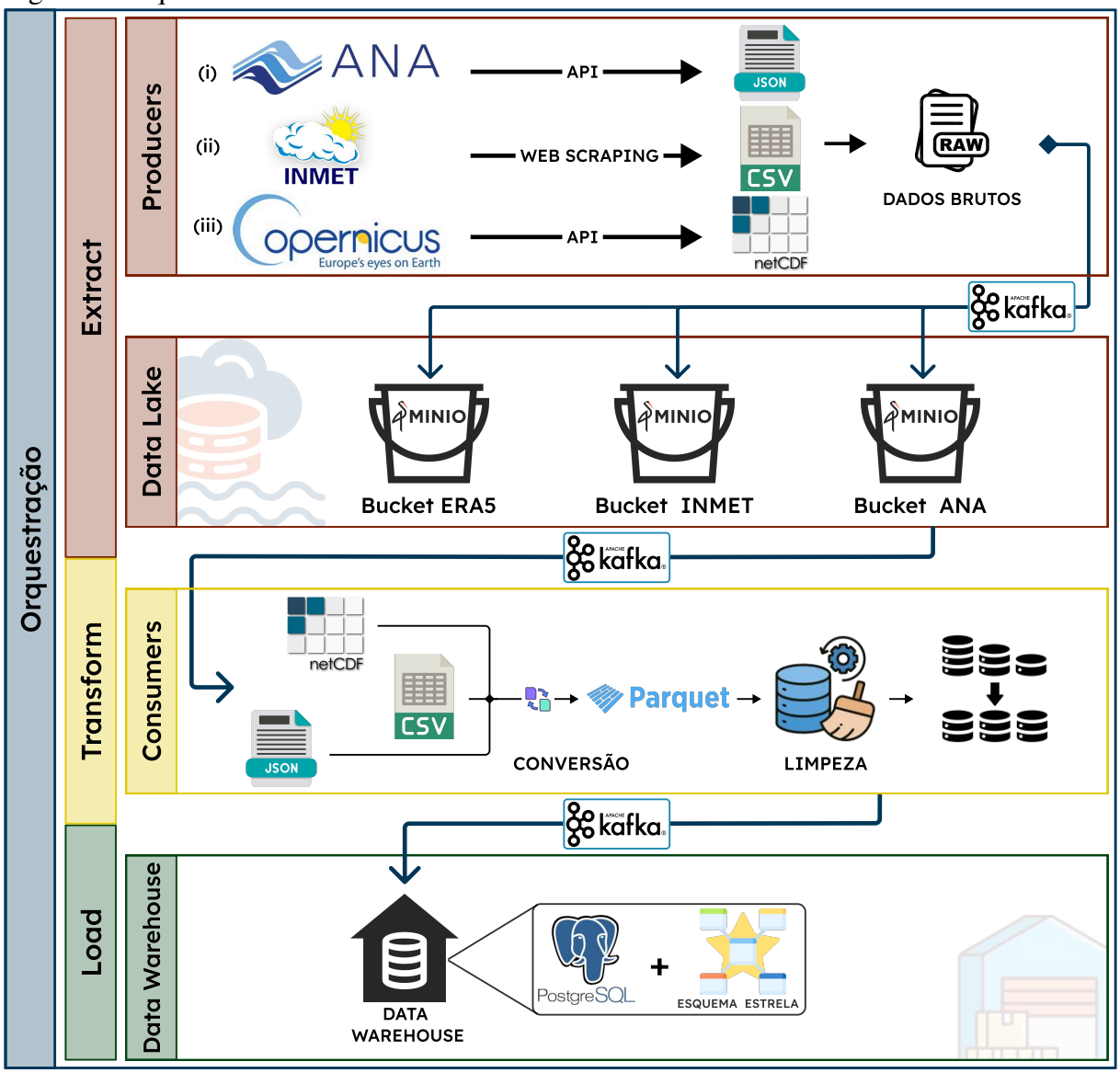
O Carcará foi implementado como um pipeline ETL assíncrono e orientado a eventos, estruturado em camadas conforme a Figura 2: *Extract*, *Transform* e *Load*. O fluxo operacional segue a sequência: (i) *Extract* coleta dados das fontes e registra o artefato bruto; (ii) um evento é publicado no *Kafka* para disparar processamento assíncrono; (iii) *Transform* consome os eventos de dados brutos, processa-os e materializa o resultado em formato estruturado; (iv) um novo evento sinaliza a disponibilidade do dado transformado; e (v) *Load* consome os eventos transformados e carrega os dados no *PostgreSQL*. Em todas as etapas, os artefatos são persistidos no *MinIO*, mantendo a separação entre armazenamento de dados brutos e transformados, enquanto o *Apache Airflow* agenda e monitora a execução diária do pipeline.

No armazenamento analítico, o *data warehouse* foi estruturado em esquema estrela, com dimensões compartilhadas (`dim_tempo`, `dim_local`, `dim_reservatorio`, `dim_localidade`, `dim_estacao_inmet`) e fatos por fonte (`fact_ana_medicao`, `fact_era5_clima`, `fact_inmet_medicao`), de forma a suportar consultas OLAP e integração entre séries meteorológicas e hidrológicas.

A coleta foi restrita ao Semiárido Brasileiro, delimitado pelas coordenadas aproximadas Norte:$-2°$, Oeste:$-46.5°$, Sul:$-10°$ e Leste:$-34.5°$, conforme o IBGE, considerando apenas áreas cobertas simultaneamente pelas três bases. A integração foi realizada exclusivamente para municípios que apresentaram, de forma concomitante, estações meteorológicas automáticas ativas do INMET, sensores *in situ* da ANA e cobertura espacial do ERA5. As séries do INMET e do ERA5 foram coletadas em resolução horária, enquanto os dados da ANA foram utilizados em resolução diária, correspondente à máxima granularidade disponibilizada.

Adicionalmente, considerou-se a defasagem de disponibilização do ERA5, cujos dados tornam-se acessíveis apenas após aproximadamente seis dias. Em virtude disso, a rotina de coleta foi parametrizada para sempre consultar o ERA5 na data $D-6$ em relação ao dia de execução do pipeline. Para manter a consistência temporal entre as fontes e garantir a mesma janela de integração, as coletas do INMET e da ANA também foram alinhadas à mesma data de referência ($D-6$).

Figura 2: Arquitetura do sistema Carcará.



Fonte: Elaborado pelo autor.

Na etapa *Transform*, o *worker* `workers/transform_worker.py` consome continuamente mensagens de `carcara.raw.fonte`; para cada evento, recupera o arquivo bruto no *MinIO*, realiza extração quando aplicável (e.g., `ZIP` do INMET) e executa o script de transformação correspondente via `subprocess.call()`, aplicando limpeza/normalização e convertendo o resultado para `Parquet`. Em seguida, o `Parquet` é enviado ao bucket `tr-fonte` e um novo evento é publicado em `carcara.tr.fonte`.

A etapa *Load* é executada pelo *worker* `workers/load_worker.py`, que consome `carcara.tr.fonte`, baixa o `Parquet` transformado do *MinIO*, conecta ao *PostgreSQL* via *SQLAlchemy* e realiza *upsert* nas dimensões e carga nas tabelas fato do esquema estrela. A execução é containerizada via *Docker Compose* e o DAG `carcara_etl_daily` é agendado para rodar diariamente às 23h, com monitoramento pelo *Airflow*.

4.3 Extract

A etapa *Extract* do fluxo ETL foi implementada por scripts em `1.extract/` acionados pelo *BashOperator* no *Apache Airflow*: para cada fonte, o extrator conecta-se à API/portal, baixa os dados do último dia ou do período configurado, salva o arquivo bruto em `downloads/`, publica um `ObjectEvent` no tópico `carcara.raw.fonte` com metadados do objeto e realiza o *upload* para o bucket `raw-fonte` no *MinIO*; a execução é considerada bem-sucedida quando o artefato é persistido no *MinIO* e o evento é publicado no *Kafka*, sendo falhas de *download*, *upload* ou publicação registradas nos logs/retorno do *Airflow*.

- **ANA (**`1.extract/ana/request_ana.py`**):** entrada via API REST; parâmetros de *início/fim* e códigos de reservatórios; saída em CSV/JSON com medições diárias, armazenada em `raw-ana` e sinalizada em `carcara.raw.ana`.
- **ERA5 (**`1.extract/era5/request_era5.py`**):** entrada via CDS; parâmetros de variáveis climáticas, período e área geográfica; saída em `NetCDF`, armazenada em `raw-era5` e sinalizada em `carcara.raw.era5`.
- **INMET (**`1.extract/inmet/request_bdmep.py`**):** entrada por *web scraping* do portal; parâmetros de códigos de estações e período; saída em `ZIP` contendo CSVs de medições horárias, armazenada em `raw-inmet` e sinalizada em `carcara.raw.inmet`.

4.4 Transform

A etapa *Transform* foi executada continuamente pelo *worker* `transform_worker.py`. O procedimento para cada mensagem consumida de `carcara.raw.{fonte}` foi: (i) baixar do *MinIO* o artefato bruto referenciado no evento; (ii) extrair artefatos quando aplicável, como `ZIP` do INMET; (iii) executar o script de transformação da fonte via `subprocess.call()`; (iv) gerar um arquivo `Parquet` com dados limpos/normalizados; (v) fazer upload do resultado para o bucket `tr-{fonte}` no *MinIO*; e (vi) publicar um novo evento em `carcara.tr.{fonte}`. A etapa foi considerada bem-sucedida quando o `Parquet` foi produzido e persistido no *MinIO*, seguido da publicação do evento transformado; falhas de consumo, download, execução do subprocesso ou upload caracterizaram erro.

- **ANA (**`2.transform/ana/transform_ana.py`**):** entrada CSV/JSON bruto; processamento com cálculo de médias e remoção de *outliers*; saída `Parquet` com colunas padronizadas (`data, reservatorio_id, volume, nivel`).
- **ERA5 (**`2.transform/era5/transform_era.py`**):** entrada `NetCDF`; conversão e agregação por hora/dia; saída `Parquet` (`data, local_id, temperatura, precipitacao`).
- **INMET (**`2.transform/inmet/transform_imet.py`**):** entrada `ZIP` com múltiplos CSVs; extração e padronização de colunas; saída `Parquet` (`data, estacao_id, temperatura, pressao`).

4.5 Load

A etapa *Load* foi executada continuamente pelo *worker* `workers/load_worker.py`, consumindo `carcara.tr.{fonte}`. Para cada evento, o procedimento foi: (i) baixar do *MinIO* o `Parquet` transformado; (ii) conectar ao *PostgreSQL* via `SQLAlchemy`; (iii) realizar *upsert*, *insert/update*, nas dimensões do esquema estrela; (iv) inserir os registros na tabela fato correspondente; e (v) confirmar a transação para garantir atomicidade. A etapa foi considerada bem-sucedida quando a transação foi concluída sem erro e os registros foram materializados nas dimensões e fatos; falhas de conexão, conflito de dados ou erro transacional caracterizaram falha

de carga.

4.6 Orquestração

A automação e o monitoramento do *pipeline* foram realizados pelo *Apache Airflow*, implantado via *Docker Compose* nos serviços *init*, *webserver*, porta 8080, e *scheduler*, com definição de DAGs e acompanhamento pela interface *web*. O DAG `carcara_etl_daily` é executado diariamente às 23h e aciona, via *BashOperator*, os scripts da etapa *Extract*. As etapas *Transform* e *Load* não são encadeadas como tarefas do DAG: elas são executadas por *workers* em loop contínuo que consomem, respectivamente, os tópicos `carcara.raw.{fonte}` e `carcara.tr.{fonte}`, processando os artefatos persistidos no MinIO e mantendo o desacoplamento do fluxo orientado a eventos. A observabilidade do processo foi realizada pelos estados de execução no *Airflow* e pelos logs dos contêineres/serviços, para identificar falhas e verificar a execução recorrente do *pipeline*.

5 EXPERIMENTAÇÃO

O sistema foi testado em ambiente de laboratório, utilizando uma máquina dedicada como servidor (Intel Core i7-9750H, NVIDIA GeForce GTX 1660 Ti 4GB GDDR5, 16GB DDR4 e SSD NVMe PCIe de 512GB), com o objetivo de verificar o funcionamento integrado dos componentes e o fluxo ponta a ponta do pipeline. Após essa etapa inicial, foram identificados ajustes necessários para assegurar a execução sem impedimentos, destacando-se, primeiramente, a implementação de uma verificação prévia de requisições pendentes no INMET, uma vez que o BDMEP impõe a restrição de apenas uma requisição ativa por vez. Adicionalmente, observou-se tempo de resposta irregular no atendimento das requisições do INMET, podendo atingir até 24 horas e exceder o *timeout* de espera configurado; como solução, adotou-se um mecanismo de verificações periódicas para detecção do retorno e disparo do *download* assim que o artefato fosse disponibilizado. Verificou-se ainda que as requisições horárias não mantinham o padrão temporal pretendido em virtude das diferenças de latência de disponibilização entre as fontes, dado que ANA e ERA5 apresentam retorno imediato enquanto o INMET pode variar; em resposta, definiu-se como estratégia de execução a submissão das requisições apenas na última hora do dia de referência, mitigando desalinhamentos e preservando a sincronização temporal. Após a incorporação desses ajustes, o sistema foi colocado em operação em 27/12/2025 e executado até 03/01/2026, período no qual todas as funcionalidades previstas foram realizadas com sucesso.

6 RESULTADOS

A validação da arquitetura foi realizada em laboratório para verificar a integração dos componentes e a execução ponta a ponta do *pipeline*. Nessa etapa, foram necessários ajustes devido às limitações do INMET, incluindo latência variável, restrição a uma requisição ativa, variação de tempo de resposta e checagens periódicas para iniciar o *download* quando o artefato fosse disponibilizado, além da submissão das requisições na última hora do dia de referência para reduzir desalinhamentos temporais entre as fontes. Após esses ajustes, o sistema entrou em operação em 27/12/2025 e permaneceu em execução até 03/01/2026.

No *Extract*, a coleta foi parametrizada para a data $D-6$, motivada pela defasagem de disponibilização do ERA5; para manter consistência temporal, as coletas do INMET e da ANA foram alinhadas à mesma referência. A integração foi restrita a municípios com cobertura concomitante das três bases, estações automáticas ativas do INMET, sensores *in situ* da ANA

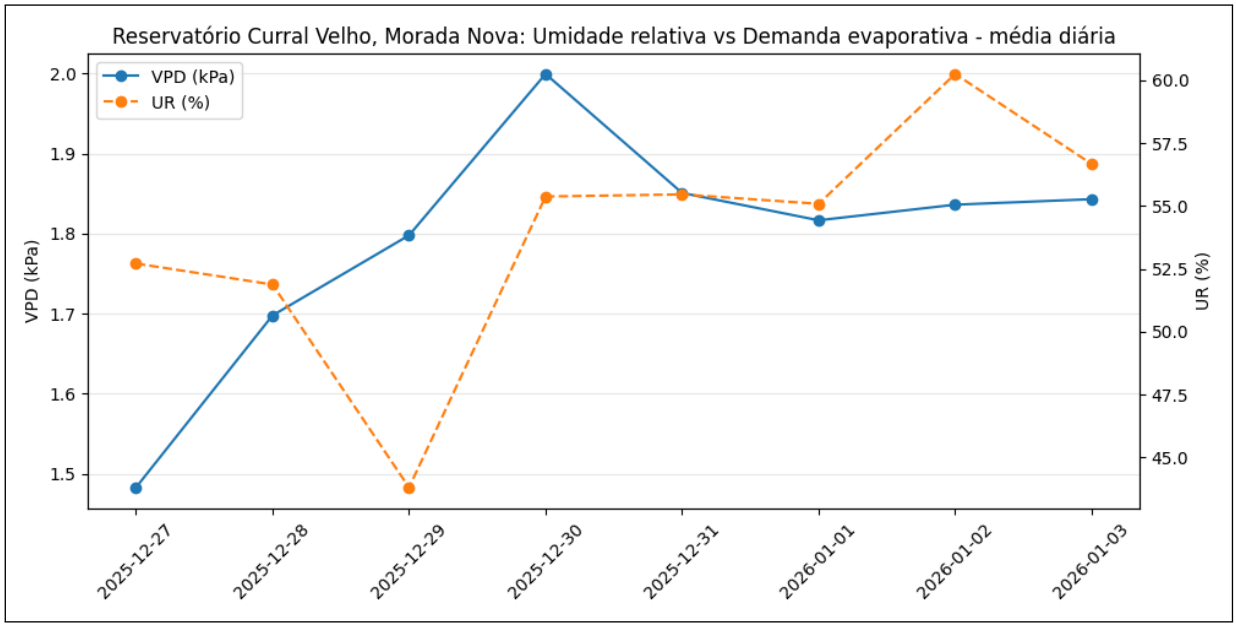
e cobertura espacial do ERA5; as séries do INMET e do ERA5 foram coletadas em resolução horária, enquanto a ANA foi utilizada em resolução diária. Operacionalmente, os extratores foram acionados pelo *BashOperator* no *Apache Airflow*; para cada fonte, o artefato bruto foi salvo, enviado ao MinIO no *bucket* `raw-<fonte>` e anunciado por um *ObjectEvent* no tópico Kafka `carcara.raw.<fonte>`.

Na etapa *Transform*, um *worker* consumiu continuamente `carcara.raw.<fonte>`; a cada evento, recuperou o artefato no MinIO, extraiu quando aplicável, executou o *script* da fonte e gravou `Parquet` limpo/normalizado no *bucket* `tr-<fonte>`, publicando então `carcara.tr.<fonte>`. Na etapa *Load*, o *worker* `workers/load_worker.py` consumiu `carcara.tr.<fonte>`, baixou o `Parquet`, conectou ao PostgreSQL via *SQLAlchemy* e realizou *upsert* nas dimensões e inserção nas tabelas fato do esquema estrela. O *Airflow* orquestrou a execução: o DAG `carcara_etl_daily` roda diariamente às 23h para acionar *Extract*, enquanto *Transform* e *Load* operam em loop contínuo via consumo dos tópicos, com monitoramento por logs. O sistema e o *dataset* resultante estão disponíveis em https://github.com/mikaelmota13/carcara.git.

6.1 Aplicações do Carcará

A partir do dataset construído neste trabalho, os dados integrados permitem diferentes análises do clima no semiárido brasileiro, combinando variáveis observadas do INMET, ANA e ERA5. A Figura 3 exemplifica uma análise da demanda evaporativa atmosférica ao relacionar a umidade relativa média diária (UR) ao déficit de pressão de vapor (VPD), em que VPD $= e_s(T) - e_a(T_d)$ e valores maiores indicam atmosfera mais seca.

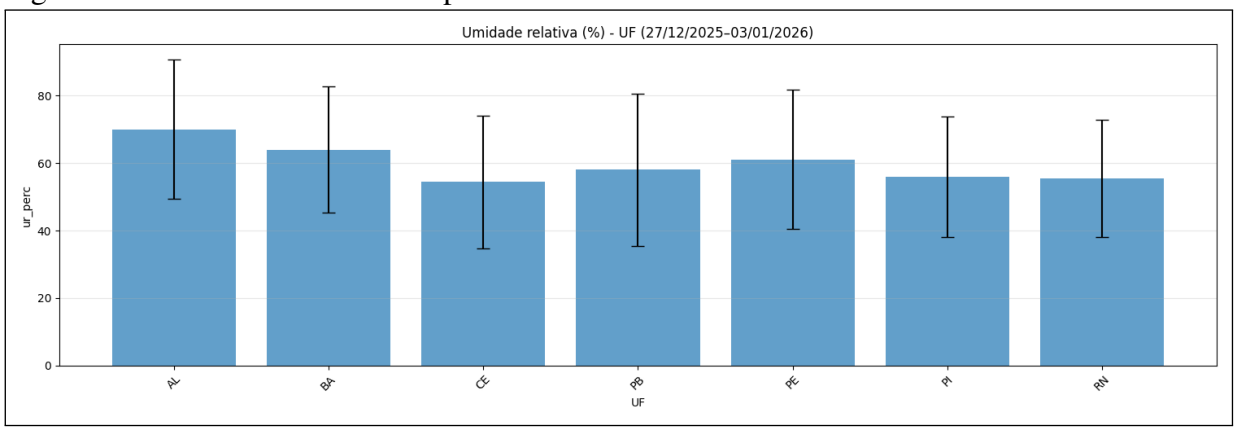
Figura 3: Relação entre umidade relativa e demanda evaporativa.



Fonte: Elaborado pelo autor.

A Figura 4 ilustra uma análise espacial comparativa ao sintetizar a UR média por UF no período, apresentando a média e ± 1 desvio-padrão como medida de variabilidade temporal e entre municípios. Para evitar viés pela quantidade de reservatórios, os dados foram previamente agregados por UF, município e data antes do cálculo; a variabilidade observada pode refletir o ciclo diurno e eventos meteorológicos mesmo em janelas curtas.

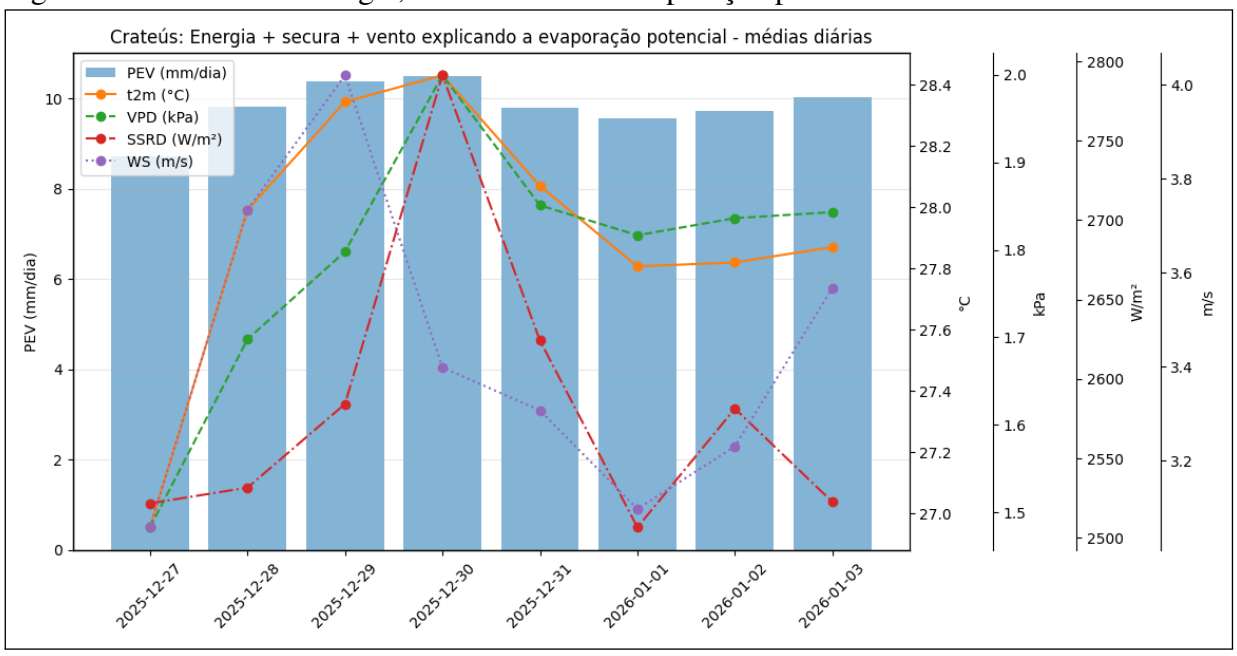
Figura 4: Umidade relativa média po UF



Fonte: Elaborado pelo autor.

A Figura 5 demonstra a possibilidade de investigar os controladores meteorológicos da evaporação potencial diária (PEV), evidenciando aumentos de PEV em dias com maior radiação solar (SSRD), maior secura do ar (VPD), temperatura (t2m) mais elevada e, em alguns casos, vento mais intenso (WS).

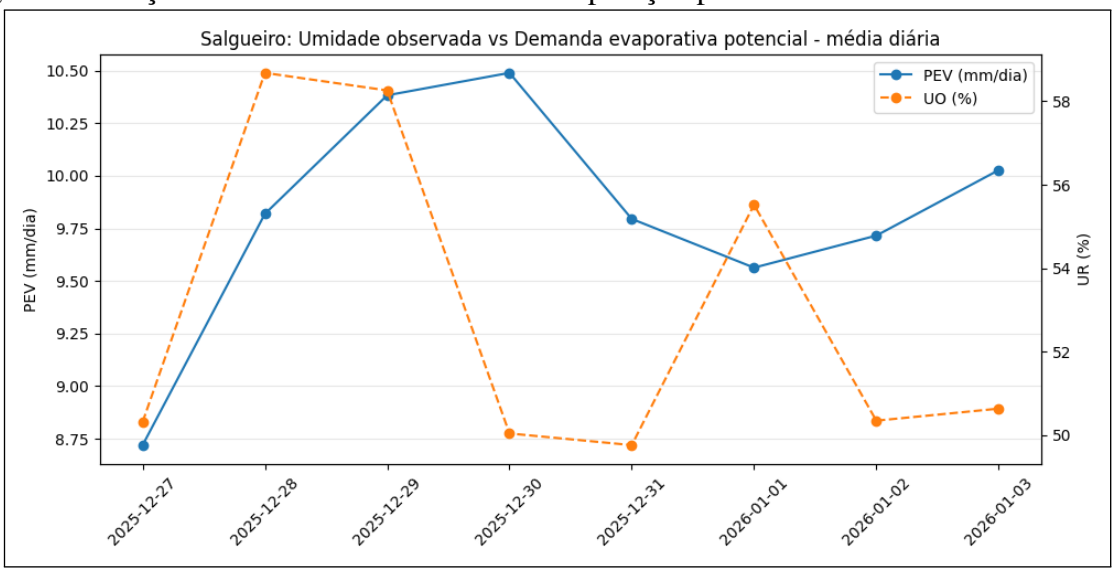
Figura 5: Influência de energia, VPD e vento na evaporação potencial.



Fonte: Elaborado pelo autor.

Por fim, a Figura 6 mostra uma análise integrada entre fontes ao comparar a PEV com a umidade observada (UO), indicando PEV mais alta em dias mais secos, com pico em 30/12 ($\sim$10,5 mm/dia), e redução da PEV quando a umidade aumenta (01/01).

Figura 6: Relação entre umidade observada e evaporação potencial



Fonte: Elaborado pelo autor.

7 CONCLUSÕES E TRABALHOS FUTUROS

Conclui-se que a arquitetura ETL *event-driven* implementada viabilizou a aquisição, padronização, integração e persistência analítica de dados meteorológicos e hidrológicos de fontes heterogêneas, com coordenação reprodutível via Apache Airflow e desacoplamento por mensageria Kafka. O fluxo ponta a ponta (MinIO em camadas e publicação de eventos `carcara.raw.<fonte>` e `carcara.tr.<fonte>`, seguido de carga no PostgreSQL) foi validado em laboratório após ajustes operacionais motivados principalmente pelas limitações do INMET, como latência variável e restrição a uma requisição ativa, além de checagens periódicas para liberação do artefato.

No período experimental, o sistema materializou um conjunto integrado com 17.472 registros em resolução horária, cobrindo 192 timestamps (27/12/2025 00:00 a 03/01/2026 23:00), 44 municípios, 8 UFs e 91 reservatórios, preservando uma grade temporal uniforme para integração. A consistência temporal foi garantida pela adoção de uma janela de referência única (D−6) alinhada à defasagem de disponibilização do ERA5, com as demais fontes ajustadas à mesma data de referência.

Quanto à completude das fontes no dataset final, observou-se cobertura integral para as variáveis ERA5 no intervalo avaliado, enquanto a ANA apresentou medições de nível/volume (*cota*, *volume* e *volumeUtil*) para 58 de 91 reservatórios (63,7% dos registros). O INMET permaneceu como principal limitante de cobertura, com dados disponíveis em 20 de 44 municípios e ao menos um campo meteorológico preenchido em 39,6% dos registros. Assim, a validação caracteriza uma prova de execução e integração ponta a ponta do *pipeline* sob restrições reais das fontes, e não uma caracterização exaustiva de cobertura e estabilidade operacional em longo prazo.

Como trabalhos futuros, propõe-se: (i) ampliar e robustecer a cobertura e integrar fontes com periodicidade próxima ao tempo real; (ii) incorporar verificações sistemáticas de qualidade como completude, consistência temporal, detecção de lacunas e outliers e observabilidade das métricas de latência por fonte, taxa de falhas e *retries*; (iii) desenvolver uma camada de disponibilização ao usuário com consulta, filtragem e visualização; e (iv) estender a cobertura espacial para além do Semiárido Brasileiro, generalizando parâmetros de extração, recortes geográficos e regras de integração para diferentes contextos territoriais.

REFERÊNCIAS

ABIRI, R.; RIZAN, N.; BALASUNDRAM, S. K.; SHAHBAZI, A. B.; ABDUL-HAMID, H. Application of digital technologies for ensuring agricultural productivity. **Heliyon**, v. 9, 2023.

AKPAN, I. J.; KOBARA, Y. M.; OWOLABI, J.; AKPAN, A. A.; OFFODILE, O. F. Conversational and generative artificial intelligence and human–chatbot interaction in education and research. **International Transactions in Operational Research**, v. 32, n. 3, p. 1251–1281, 2025. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/itor.13522.

AMIRA, Y. H. e Bhagawat Rimal e A. Tiwary e A. Utilizando inteligência artificial e fusão de dados para monitoramento ambiental: uma revisão e perspectivas futuras. **Inf. Fusion**, v. 86-87, p. 44–75, 2022.

ANA. **Portal da Agência Nacional de Águas e Saneamento Básico (ANA)**. 2025. Acesso em: 11 nov. 2025. Disponível em: https://www.gov.br/ana/pt-br.

Apache Software Foundation. **Apache Kafka**. 2024. https://kafka.apache.org/. Acessado em: 11 nov. 2025.

BARBOSA, H. Understanding the rapid increase in drought stress and its connections with climate desertification since the early 1990s over the brazilian semi-arid region. **Journal of Arid Environments**, Elsevier, v. 222, p. 105142, 2024.

BHARADIYA, J. P.; TZENIOS, N. T.; REDDY, M. Predicting crop yield using deep learning and remote sensing. **Journal of Engineering Research and Reports**, v. 24, n. 12, p. 29–44, 2023.

CEPEA, C. de Estudos Avançados em E. A. MERCADO DE TRABALHO/CEPEA: Agronegócio emprega 28,2 milhões de pessoas em 2024 e representa 26% das ocupações do País. 2025. Disponível em: https://l1nq.com/Mnqsm.

CEPEA, C. de Estudos Avançados em E. A. PIB-Agro/CEPEA: Desempenho do 4º trimestre reverte tendência de queda anual, e PIB do agronegócio avança 1,81% em 2024. 2025. Disponível em: https://sl1nk.com/ko48H.

CHANDRA, P.; GUPTA, M. K. Comprehensive survey on data warehousing research. **International Journal of Information Technology**, Springer, v. 10, n. 2, p. 217–224, 2018.

CHEN, J.; LV, N. Using big data and machine learning to optimize agricultural resource allocation and ecological and economic benefit evaluation. **International Journal of Agricultural and Environmental Information Systems**, 2025.

CHERRADI, M.; BOUHAFER, F.; HADDADI, A. E. Data lake governance using ibm-watson knowledge catalog. **Scientific African**, 2023.

CONTINI, E.; ARAGÃO, A. O agro brasileiro alimenta 800 milhões de pessoas. **Brasília: Embrapa**, 2021.

Câmara dos Deputados. **Deputados defendem fortalecimento da agricultura para produção de mais alimentos**. 2023. Disponível em: https://www.camara.leg.br/noticias/937752-deputados-defendem-fortalecimento-da-agricultura-para-producao-de-mais-alimentos/.

DAVENPORT, T. **Big data at work: dispelling the myths, uncovering the opportunities**. [*S. l.*]: Harvard Business Review Press, 2014.

DHANDA, M.; ROGERS, B. A.; HALL, S.; DEKONINCK, E.; DHOKIA, V. Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0. **Robotics and Computer-Integrated Manufacturing**, v. 93, p. 102937, 2025. ISSN 0736-5845. Disponível em: https://www.sciencedirect.com/science/article/pii/S0736584524002242.

DIXON, J. **Pentaho, Hadoop, and Data Lakes**. 2010. Acesso em: 10 nov. 2025. Disponível em: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

D'ACQUISTO, G.; DOMINGO-FERRER, J.; KIKIRAS, P.; TORRA, V.; MONTJOYE, Y. de; BOURKA, A. Privacy by design in big data. **An overview of privacy enhancing technologies in the era of big data analytics. European Union Agency for Network and Information Security**, v. 79, 2015.

EL-BASTAWISSY, S. E.-S. e Abdeltawab M. Hendawi e A. Um modelo proposto para processos etl de data warehouse. **J. King Saud Univ. Comput. Inf. Sci.**, v. 23, p. 91–104, 2011.

FIDALGO, F. F. e Robson do N. Uma análise de desempenho de bancos de dados em nuvem híbridos e colunares para projeto de esquema eficiente em data warehouse distribuído como serviço. **Data**, v. 9, p. 99, 2024.

FORTINI, R. M.; BRAGA, M. J. **Um novo retrato da agricultura familiar do semiárido nordestino brasileiro: a partir dos dados do censo agropecuário 2017**. [*S. l.*]: Universidade Federal de Viçosa, 2020.

FRAGOSO, E. J. N.; COELHO, P.; SOUZA, P.; NETO, A. F.; SANTIAGO, A. M. dos S.; PACHECO, C. S. G. R.; MELO, R. Agroecology and family farming: A perspective of sustainability in the brazilian semiarid region. **International Journal of Advanced Engineering Research and Science**, 2020.

GARG, D.; ALAM, M. A sensor data acquisition system for smart agriculture. **SN Computer Science**, Springer, v. 4, n. 5, p. 667, 2023.

GEWALI, U. B.; MONTEIRO, S. T.; SABER, E. Machine learning based hyperspectral image analysis: a survey. **arXiv preprint arXiv:1802.08701**, 2018.

HAI, R.; KOUTRAS, C.; QUIX, C.; JARKE, M. Data lakes: A survey of functions and systems. **IEEE Transactions on Knowledge and Data Engineering**, v. 35, p. 12571–12590, 2021.

HAN, S.-W. **Three-tier architecture for sentinel applications and tools: separating presentation from functionality**. Dissertação (Mestrado) – University of Florida, 1997.

HIMESH, S.; RAO, E. P.; GOUDA, K.; RAMESH, K.; RAKESH, V.; MOHAPATRA, G.; RAO, B. K.; SAHOO, S. K.; AJILESH, P. Digital revolution and big data: a new revolution in agriculture. **CABI Reviews**, CABI Wallingford UK, n. 2018, p. 1–7, 2018.

IBM. **MinIO para IBM Cloud Private**. 2023. Disponível em: https://www.ibm.com/docs/pt-br/cloud-private/3.2.x?topic=private-minio.

IBM. **O que é Data Warehouse?** 2025. https://www.ibm.com/br-pt/think/topics/data-warehouse. Acesso em: 29 out. 2025.

INMET. **Portal do Instituto Nacional de Meteorologia (INMET)**. 2025. https://portal.inmet.gov.br/. Disponível em: https://portal.inmet.gov.br/. Acesso em: 11 nov. 2025.

INMON, W. H. **Building the data warehouse**. [*S. l.*]: John wiley & sons, 2005.

INSA. **Semiárido Brasileiro**. 2023. Disponível em: https://www.gov.br/insa/pt-br/semiarido-brasileiro.

JIANG, Z. **The impact of sustainable management systems on the sustainable development of water resources in agriculture and the reduction of agricultural carbon footprint**. 2025. 135504D - 135504D-8 p.

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.

KAMILARIS, A.; PRENAFETA-BOLDÚ, F. X. Deep learning in agriculture: A survey. **Computers and electronics in agriculture**, Elsevier, v. 147, p. 70–90, 2018.

KHAN, B.; JAN, S.; KHAN, W.; CHUGHTAI, M. I. An overview of etl techniques, tools, processes and evaluations in data warehousing. **Journal on Big Data**, 2024.

KIMBALL, R.; CASERTA, J. **The data warehouse ETL toolkit**. [*S. l.*]: John Wiley & Sons, 2004.

KIMBALL, R.; ROSS, M. **The data warehouse toolkit: The definitive guide to dimensional modeling**. [*S. l.*]: John Wiley & Sons, 2013.

KNEZEVIC, K. B. e I. Big data na alimentação e agricultura. **Big Data Society**, v. 3, 2016.

KUNE, R.; KONUGURTHI, P. K.; AGARWAL, A.; CHILLARIGE, R. R.; BUYYA, R. The anatomy of big data computing. **Software: Practice and Experience**, v. 46, n. 1, p. 79–105, 2016. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2374.

LI, S.; DRAGICEVIC, S.; CASTRO, F. A.; SESTER, M.; WINTER, S.; COLTEKIN, A.; PETTIT, C.; JIANG, B.; HAWORTH, J.; STEIN, A.; CHENG, T. Geospatial big data handling theory and methods: A review and research challenges. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 115, p. 119–133, 2016. ISSN 0924-2716. Theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'. Disponível em: https://www.sciencedirect.com/science/article/pii/S0924271615002439.

MADERA, C.; LAURENT, A. The next information architecture evolution: the data lake wave. In: **Proceedings of the 8th international conference on management of digital ecosystems**. [*S. l.*: *s. n.*], 2016. p. 174–180.

MARCH, S.; HEVNER, A. Integrated decision support systems: A data warehousing perspective. **Decis. Support Syst.**, v. 43, p. 1031–1043, 2007.

MARENGO, J.; CUNHA, A. P.; NOBRE, C.; NETO, G. R. R.; MAGALHãES, A.; TORRES, R.; SAMPAIO, G.; ALEXANDRE, F.; ALVES, L.; CUARTAS, L. A.; DEUSDARá, K. R. L.; ALVALá, R. Assessing drought in the drylands of northeast brazil under regional warming exceeding 4 °c. **Natural Hazards**, p. 1–23, 2020.

MARENGO, J. A.; TORRES, R. R.; ALVES, L. M. Drought in northeast brazil—past, present, and future. **Theoretical and Applied Climatology**, Springer, v. 129, n. 3, p. 1189–1200, 2017.

MARINELLO, F.; ZOU, X.; LIU, Z.; ZHU, X.; ZHANG, W.; QIAN, Y.; LI, Y.; KARUNATHILAKE, E.; LE, A. T.; HEO, S.; CHUNG, Y.; MANSOOR, S. The path to smart farming: Innovations and opportunities in precision agriculture. **Agriculture**, 2023.

MEGDICHE, I.; RAVAT, F.; ZHAO, Y. Metadata management on data processing in data lakes. p. 553–562, 2021.

MENA, F.; PATHAK, D.; NAJJAR, H.; SANCHEZ, C.; HELBER, P.; BISCHKE, B.; HABELITZ, P.; MIRANDA, M.; SIDDAMSETTY, J.; NUSKE, M. *et al.* Adaptive fusion of multi-view remote sensing data for optimal sub-field crop yield prediction. **arXiv preprint arXiv:2401.11844**, 2024.

MINIO, D. **Hybrid and Multi-Cloud Object Storage Solutions**. 2025. Página institucional sobre soluções híbridas e multicloud. Disponível em: https://www.min.io/solutions/hybrid-multi-cloud-storage.

MUNAWAR. Qualidade de dados baseada em extração, transformação e carregamento (etl) para desenvolvimento de data warehouse. **2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)**, v. 1, p. 373–378, 2021.

MUSANASE, C.; VODACEK, A.; HANYURWIMFURA, D.; UWITONZE, A.; KABANDANA, I. Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices. **Agriculture**, 2023.

MUÑOZ, A. C. e Sebastián Pardo e S. Sepúlveda e L. Desafios para o uso de aprendizado de máquina em big data agrícola: uma revisão sistemática da literatura. **Agronomy**, 2022.

MUñOZ-SABATER, J.; DUTRA, E.; AGUSTí-PANAREDA, A.; ALBERGEL, C.; ARDUINI, G.; BALSAMO, G.; BOUSSETTA, S.; CHOULGA, M.; HARRIGAN, S.; HERSBACH, H.; AUTHORS other. Era5-land: A state-of-the-art global reanalysis dataset for land applications. **Earth System Science Data**, Copernicus Publications, v. 13, p. 4349–4383, 2021. Disponível em: https://doi.org/10.5194/essd-13-4349-2021.

NAMBIAR, A.; MUNDRA, D. An overview of data warehouse and data lake in modern enterprise data management. **Big Data and Cognitive Computing**, v. 6, n. 4, 2022. ISSN 2504-2289. Disponível em: https://www.mdpi.com/2504-2289/6/4/132.

NEHA, R. V. e S. Uma breve nota sobre data warehouse. **ASEAN Journal of Psychiatry**, 2024.

NEUMEYER, X.; SANTOS, S.; MORRIS, M. H. Poverty entrepreneurs and technology. **2018 IEEE Technology and Engineering Management Conference (TEMSCON)**, p. 1–5, 2018.

PALOPOLI, L.; PONTIERI, L.; TERRACINA, G.; URSINO, D. A novel three-level architecture for large data warehouses. **J. Syst. Archit.**, v. 47, p. 937–958, 2002.

PECUÁRIA, M. da Agricultura e. **Exportações do agronegócio ultrapassam US$ 153 bilhões no acumulado de 2024**. 2024. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/noticias/2024/exportacoes-do-agronegocio-ultrapassam-us-153-bilhoes-no-acumulado-de-2024.

PELLEGRINA, H. S. Trade, productivity, and the spatial organization of agriculture: Evidence from brazil. **Journal of Development Economics**, 2022.

PostgreSQL Global Development Group. **PostgreSQL A relational database management system**. 2024. Acesso em: 10 nov. 2025. Disponível em: https://www.postgresql.org/.

QAISER, A.; FAROOQ, M. U.; MUSTAFA, S. M. N.; ABRAR, N. Comparative analysis of etl tools in big data analytics. **Pakistan Journal of Engineering and Technology**, 2023.

QIANG, S.; LIU, L. Research on the design of a new data warehouse system. **2009 2nd IEEE International Conference on Computer Science and Information Technology**, p. 462–465, 2009.

RAHMAN, M. B.; CHAKMA, J. D.; MOMIN, A.; ISLAM, S.; UDDIN, M. A.; ISLAM, M. A.; ARYAL, S. Smart crop cultivation system using automated agriculture monitoring environment in the context of bangladesh agriculture. **Sensors**, v. 23, n. 20, 2023. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/23/20/8472.

RAHMAN, M. B.; CHAKMA, J. D.; MOMIN, A.; ISLAM, S.; UDDIN, M. A.; ISLAM, M. A.; ARYAL, S. Smart crop cultivation system using automated agriculture monitoring environment in the context of bangladesh agriculture. **Sensors**, MDPI, v. 23, n. 20, p. 8472, 2023.

RAJIANI, I.; KOT, S.; MICHAŁEK, J.; RIANA, I. G. Barriers to technology innovation among nascent entrepreneurs in deprived areas. **Problems and Perspectives in Management**, 2023.

RAVAT, F.; ZHAO, Y. Data lakes: Trends and perspectives. In: SPRINGER. **International Conference on Database and Expert Systems Applications**. [*S. l.*], 2019. p. 304–313.

RAZA, U.; KULKARNI, P.; SOORIYABANDARA, M. Low power wide area networks: An overview. **ieee communications surveys & tutorials**, IEEE, v. 19, n. 2, p. 855–873, 2017.

RELLA, B. P. R. Análise comparativa de data lakes e data warehouses para aprendizado de máquina. **International Journal For Multidisciplinary Research**, 2025.

REPOSITORYMINIO. **MinIO on GitHub**. 2025. Repositório oficial do projeto MinIO. Disponível em: https://github.com/minio/minio.

ROPER, J.; GARCíA, J. F.; TSUTSUI, H. Emerging technologies for monitoring plant health in vivo. **ACS Omega**, v. 6, p. 5101 – 5107, 2021.

SARANYA, G.; KUMARAN, K.; PRATHEEP, V.; DHANUSH, S. Developing scalable and fault tolerant distributed architecture for big unstructured data systems. In: IEEE. **2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)**. [*S. l.*], 2025. p. 761–768.

SAVEINCLOUD. **MinIO: o que é e como funciona o serviço de armazenamento em nuvem**. 2023. Disponível em: https://saveincloud.com/pt/blog/armazenamento/minio/.

SAWADOGO, P.; DARMONT, J. On data lake architectures and metadata management. **Journal of Intelligent Information Systems**, Springer, v. 56, n. 1, p. 97–120, 2021.

SAWADOGO, P. N.; DARMONT, J. On data lake architectures and metadata management. **Journal of Intelligent Information Systems**, v. 56, p. 97 – 120, 2020.

SHARMA, K.; SHIVANDU, S. kumar. Integrating artificial intelligence and internet of things (iot) for enhanced crop monitoring and management in precision agriculture. **Sensors International**, 2024.

SILVA, J. R. d.; MEDEIROS, F. B. d.; MOURA, F. M. S. d.; BESSA, W. d. S.; BEZERRA, E. L. M. Uso das tecnologias de informação e comunicação no curso de medicina da ufrn. **Revista Brasileira de Educação Médica**, Associação Brasileira de Educação Médica, v. 39, n. 4, p. 537–541, Oct 2015. ISSN 0100-5502. Disponível em: https://doi.org/10.1590/1981-52712015v39n4e02562014.

STEPHEN, A. *et al.* Using open remote sensing data to build an agriculture big data system. **Turkish Journal of Computer and Mathematics Education (TURCOMAT)**, v. 12, n. 2, p. 429–436, 2021.

TAURION, C. **Big data**. [*S. l.*]: Brasport, 2013.

TISIN, S. R.; OTHMAN, N. Entrepreneurs' challenges in mastering digital technology skills. **Advanced International Journal of Business, Entrepreneurship and SMEs**, 2024.

VASSILIADIS, P. A survey of extract–transform–load technology. **International Journal of Data Warehousing and Mining (IJDWM)**, IGI Global Scientific Publishing, v. 5, n. 3, p. 1–27, 2009.

VASSILIADIS, P.; SIMITSIS, A.; SKIADOPOULOS, S. Modeling etl activities as graphs. In: **DMDW**. [*S. l.*: *s. n.*], 2002. v. 58, p. 52–61.

VIANA, C. M.; FREIRE, D.; ABRANTES, P.; ROCHA, J.; PEREIRA, P. Agricultural land systems importance for supporting food security and sustainable development goals: A systematic review. **The Science of the total environment**, p. 150718, 2021.

WARNARS, H. Star schema design for concept hierarchy in attribute oriented induction. **Internetworking Indonesia Journal**, v. 2, n. 2, p. 33–39, 2010.

WEI, Y.; HAN, C.; YU, Z. An environment safety monitoring system for agricultural production based on artificial intelligence, cloud computing and big data networks. **Journal of Cloud Computing**, Springer, v. 12, n. 1, p. 83, 2023.

WIJAYA, W.; WIRATAMA, J. *et al.* The implementation of data warehouse and star schema for optimizing property business decision making. **G-Tech: Jurnal Teknologi Terapan**, v. 8, n. 2, p. 1242–1250, 2024.

WOLANIN, A.; CAMPS-VALLS, G.; GÓMEZ-CHOVA, L.; MATEO-GARCÍA, G.; TOL, C. V. D.; ZHANG, Y.; GUANTER, L. Estimating crop primary productivity with sentinel-2 and landsat 8 using machine learning methods trained with radiative transfer simulations. **Remote sensing of environment**, Elsevier, v. 225, p. 441–457, 2019.

WREMBEL, R. Desempenho de data warehouse: Técnicas e estruturas de dados selecionadas. p. 27–62, 2011.

ZHU, G.; QIN, Z.; XIONG, H.; KUMARI, S.; ALENAZI, M. J.; GUO, Y.; CHEN, C.-M. Open-world multi-modal machine learning decision model based on uncertain data analysis for fetal heart diagnosis. **Information Sciences**, v. 701, p. 121868, 2025. ISSN 0020-0255. Disponível em: https://www.sciencedirect.com/science/article/pii/S0020025524017821.