KELLY CHRISTINE ALVARENGA DE CASTRO

SOLUÇÃO DE BIG DATA PARA O MONITORAMENTO E DETECÇÃO DE FALHAS EM POÇOS DE PETRÓLEO UTILIZANDO APRENDIZADO DE MÁQUINA.

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.

SÃO PAULO

2024

KELLY CHRISTINE ALVARENGA DE CASTRO

SOLUÇÃO DE BIG DATA PARA O MONITORAMENTO E DETECÇÃO DE FALHAS EM POÇOS DE PETRÓLEO UTILIZANDO APRENDIZADO DE MÁQUINA.

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.

Área de concentração: Tecnologia da Informação – Big Data

Orientador: Prof. MEE Jonas Santiago de Oliveira

SÃO PAULO

2024

Catalogação-na-publicação

Castro, Kelly
SOLUÇÃO DE BIG DATA PARA O MONITORAMENTO E DETECÇÃO DE FALHAS EM POÇOS DE PETRÓLEO UTILIZANDO APRENDIZADO DE MÁQUINA. / K. Castro -- São Paulo, 2024.
59 p.

Monografia (Especialização em Engenharia de Dados e Big Data) - Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.Big Data 2.Detecção de falhas 3.Machine Learning 4.Métodos Ensemble 5.Computação em Nuvem I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

AGRADECIMENTOS

Durante o período de execução deste trabalho, eu perdi meu pai, Seu Vitório Paulo de Castro, a quem agradeço profundamente. Foi seu trabalho diário em uma padaria que proporcionou a melhor formação que eu poderia ter adquirido enquanto criança, formação essa que me ensinou a usar a educação para perseguir meus sonhos.

Agradeço especialmente ao Dr. Afrânio Melo, Cientista de Dados no Centro de Pesquisas, Desenvolvimento e Inovação Leopoldo Américo Miguez de Mello - CENPES/PETROBRAS, por suas mentorias e pela generosidade em compartilhar seu conhecimento e explicar detalhadamente os projetos BibMon e 3W. Foi incrível ter a colaboração do próprio autor de alguns dos artigos que uso como referência neste trabalho.

Agradeço também ao professor MEE Jonas Santiago, que, além de orientador, foi um verdadeiro mentor em minha trajetória nesta especialização na Poli USP. Desde o primeiro dia em sala de aula até a consolidação de meu aprendizado nesta monografia. Agradeço pelo acolhimento, direcionamento e pela dedicação incansável em me alertar e apontar caminhos que antes eu não enxergava. Se hoje tenho orgulho deste trabalho, é graças ao professor Jonas!

Agradeço a todos os professores do Programa de Educação Continuada da Poli USP, que me acolheram como bolsista da instituição e, em especial, à Profa. Dra. Solange Nice Alves de Souza, por todas as conversas e novos horizontes que me apresentou.

Agradeço aos meus colegas de curso, em especial ao Nikolas Gomes de Sá, à Diane Santos e ao Tomás Carvalho,que foram sempre muito generosos em compartilhar conhecimento e descobertas.

Agradeço a minha primeira gestora na indústria do petróleo, Isabelly Perico, que me ofereceu a primeira oportunidade profissional no mundo dos dados.

Dedico este trabalho à minha esposa, Maira Godoy de Carvalho, minha maior incentivadora e inspiração para ser sempre melhor a cada dia. A ela dedico e ofereço minha vida e todos os frutos do trabalho que realizamos diariamente, juntas.

CURSO ENGENHARIA DE DADOS E BIG DATA

Coord.: Prof. Solange Nice Alves de Souza
Vice-Coord.: Prof. Pedro Luiz Pizzigatti Corrêa

Perspectivas profissionais alcançadas com o curso:

Quando iniciei a especialização em Engenharia de Dados e Big Data pelo Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, havia recém ingressado em um programa de estágio em Transformação Digital, em uma indústria global do setor de petróleo e gás. Em meio a uma transição de carreira, alinhava meu bacharelado em Engenharia de Software com o objetivo de atuar em uma companhia internacional. Durante a especialização, tive a oportunidade de desenvolver projetos em parceria com a empresa, envolvendo professores e colegas na resolução de problemas reais da indústria, o que me posicionou positivamente perante meus gestores. Fui promovida antes mesmo de apresentar a monografia final e, atualmente, atuo no time global de Dados e Digitalização. Além disso, ganhei visibilidade junto a clientes e parceiros ao contribuir com um projeto Open Source de relevância para o setor.

RESUMO

Este projeto descreve a estruturação de uma instância de *Big Data* para o monitoramento e a detecção preditiva de anomalias em poços de petróleo utilizando técnicas de aprendizado de máquina. Ao integrar o *dataset* 3W e a biblioteca BibMon, de autoria da Petrobras, em uma instanciação baseada na Arquitetura de Referência NIST para *Big Data*, o sistema facilita a análise de grandes volumes de dados e aprimora a identificação precoce de falhas operacionais por meio do uso de métodos de *Ensemble*, que permitem a operação coordenada de múltiplos modelos de aprendizado de máquina. A instanciação utiliza os serviços de computação em nuvem da *Amazon Web Services* para processar, armazenar e analisar os dados de forma escalável e eficiente. Este projeto contribui para a otimização das operações na indústria de petróleo e gás, promovendo a inovação aberta e estabelecendo uma base sólida para pesquisas futuras na área.

Palavras-chave: Big Data, Aprendizado de Máquina, Métodos de Ensemble, Poços de Petróleo, Detecção de Anomalias, Computação em Nuvem, Monitoramento de Processos.

ABSTRACT

This project describes the structuring of a Big Data instance for monitoring and predictive anomaly detection in oil wells using machine learning techniques. By integrating the 3W dataset and the BibMon library, authored by Petrobras, into an instantiation based on the NIST Big Data Reference Architecture, the system facilitates the analysis of large datasets and enhances the early identification of operational failures by employing Ensemble Methods that enable the coordinated operation of multiple machine learning models. The instantiation leverages Amazon Web Services' cloud computing services to process, store, and analyze data in a scalable and efficient manner. This project contributes to the optimization of operations in the oil and gas industry, fostering open innovation and establishing a solid foundation for future research in the field.

Keywords: Big Data, Machine Learning, Ensemble Methods, Oil Wells, Anomaly Detection, Cloud Computing, Process Monitoring.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

AMV - Annulus Master Valve (Válvula Mestre do Anular)

AWS - Amazon Web Services

AWV - Annulus Wing Valve (Válvula de Asa do Anular)

BibMon - Biblioteca de Monitoramento de Processos

CENPES - Centro de Pesquisas, Desenvolvimento e Inovação Leopoldo Américo Miguez de Mello

DAG - Directed Acyclic Graph (Grafo Acíclico Direcionado)

DHSV - Downhole Safety Valve (Válvula de Segurança de Fundo de Poço)

EMR - Elastic MapReduce

ESN - Echo State Networks (Redes Neurais de Estado de Eco)

FDR - Fault Detection Rate (Taxa de Detecção de Falhas)

GLCK - Gas Lift Choke (Choke de Gás Lift)

IAM - Identity and Access Management (Gerenciamento de Identidade e Acesso)

LSTM - Long Short-Term Memory (Memória de Longo Curto Prazo)

MWAA - Managed Workflows for Apache Airflow

NBD-PWG - NIST Big Data Public Working Group

NBDRA - NIST Big Data Reference Architecture

NIST - National Institute of Standards and Technology

OTC - Offshore Technology Conference

PCA - Principal Component Analysis (Análise de Componentes Principais)

PCK - Production Choke (Choke de Produção)

PDG - Permanent Downhole Gauge (Medidor Permanente de Fundo de Poço)

PETROBRAS - Petróleo Brasileiro S.A.

PMV - Production Master Valve (Válvula Mestre de Produção)

PWV - Production Wing Valve (Válvula de Asa de Produção)

PXO - Pig Crossover Valve (Válvula de Crossover de Pig)

RNN - Recurrent Neural Network (Rede Neural Recorrente)

S3 - Simple Storage Service

SDV - Shutdown Valve (Válvula de Shutdown)

SPE - Squared Prediction Error (Erro Quadrático de Predição)

SP - Service Pump (Bomba de Serviço)

TPT - Transdutor de Temperatura e Pressão

VPC - Virtual Private Cloud

SUMÁRIO

1 INTRODUÇÃO

No contexto da indústria de óleo e gás, particularmente nas operações de *upstream* (exploração e produção), o monitoramento de processos industriais é crucial para garantir a segurança operacional, a viabilidade econômica e a redução de riscos ambientais (MELO et al., 2024). A identificação e mitigação precoces de eventos indesejáveis em poços de petróleo são essenciais para evitar perdas financeiras e minimizar impactos, especialmente em operações *offshore* (em alto-mar), onde a complexidade e a criticidade dos sistemas são acentuadas.

Nesse sentido, a transformação digital em curso tem impactado significativamente os modelos de negócios em diversos setores, criando oportunidades para geração de valor no movimento conhecido como Indústria 4.0 (ANZAI et al., 2023). A aplicação de tecnologias de Inteligência Artificial (IA) tem se expandido, influenciando a eficiência operacional e a otimização em indústrias de processo. A indústria de petróleo e gás tem seguido essa tendência, implementando a digitalização para enfrentar desafios e resolver problemas complexos.

A integração de *Big Data* e Aprendizado de Máquina (*Machine Learning)* surge como uma abordagem poderosa para lidar com grandes volumes, variedade e velocidade dos dados gerados por esses sistemas, permitindo a detecção de falhas por meio de padrões complexos e precisos (ANZAI et al., 2023). A análise avançada de dados possibilita examinar as interações entre múltiplas variáveis e identificar anomalias preditivas de falhas, aprimorando a tomada de decisão e a eficiência operacional.

Este trabalho propõe a estruturação de uma instância *Big Data* para o monitoramento e detecção de anomalias em poços de petróleo. Ao conceber um sistema robusto e escalável capaz de detectar precocemente anomalias que podem gerar falhas, este trabalho contribui para a otimização da produção, redução de custos operacionais e minimização de impactos ambientais. Além disso, avança a aplicação de tecnologias de *Big Data* e *Machine Learning* no setor, demonstrando o potencial de tais tecnologias para enfrentar desafios complexos e aprimorar a eficiência operacional.

1.1 Objetivo Geral

O objetivo deste trabalho é estruturar uma instância de *Big Data* que utilize métodos de *Ensemble* para o monitoramento e a detecção preditiva de anomalias em poços de petróleo, integrando múltiplos modelos de aprendizado de máquina para aprimorar a precisão e a robustez das previsões.

1.2 Objetivos Específicos

Para a elaboração deste trabalho, foram considerados os seguintes objetivos específicos:

1. Coletar e preparar os dados via integração com o *dataset* 3W da Petrobras para análise.
2. Estruturar uma instância de *Big Data* para processamento de dados utilizando computação em nuvem.
3. Integrar a biblioteca BibMon da Petrobras à solução para otimizar o processamento dos dados e a implementação de algoritimos de aprendizado de máquina.

1.3 Justificativa

A identificação e mitigação precoces de eventos indesejáveis em poços de petróleo são fundamentais para prevenir perdas de produção, reduzir custos de manutenção, evitar acidentes ambientais e proteger vidas humanas (PETROBRAS, 2022). Falhas não detectadas em tempo hábil podem ocasionar interrupções na produção, custos elevados de reparo e impactos ambientais significativos.

> "No segmento de *upstream*, especialmente no Brasil, o poço de petróleo é um dos focos centrais de pesquisa, pois atua como o canal que conecta a superfície ao reservatório subterrâneo. Garantir a integridade e o funcionamento eficiente dos poços é essencial para o sucesso das operações de exploração e produção.” (D'Almeida et al., 2022)

A instânciação em ambiente de computação em nuvem proposta, busca aprimorar e integrar as tecnologias de *Big Data* ao processo atual de monitoramento e detecção de anomalias, utilizando o potencial do *Big Data* para criar um ambiente favorável à aplicação de métodos de Ensemble no aprendizado de máquina para analisar grandes volumes de dados de forma eficiente e em tempo real.

Segundo D'Almeida et al. (2022, p. 5555), "essas inovações permitem a detecção e previsão de falhas, evitando ou minimizando problemas que podem resultar em altos custos ou até na perda total do poço." Essa afirmação destaca a relevância de implementar soluções tecnológicas avançadas para superar os desafios da indústria petrolífera e garantir maior eficiência operacional e sustentabilidade.

A integração da instância criada com a biblioteca BibMon da Petrobras, uma biblioteca de código aberto para monitoramento de processos, sensores virtuais e diagnóstico de falhas, contribui para o entendimento do processamento de dados no contexto dos poços de petróleo.

Ao utilizar a biblioteca BibMon aplicando técnicas de *Big Data Analytics*, ampliamos sua capacidade de processar e analisar grandes volumes de dados, tornando-a mais eficaz para aplicações em contextos como o de monitoramento de poços de petróleo. Isso não apenas beneficia a indústria de petróleo e gás, mas também fortalece a comunidade global que busca soluções eficientes e acessíveis para o monitoramento e diagnóstico de sistemas complexos.

1.4 Metodologia

Para alcançar os objetivos propostos neste trabalho, foi elaborada uma metodologia estruturada e ágil, distribuída em etapas sequenciais e incrementais, que combinam pesquisa bibliográfica, estudo de documentação e desenvolvimento da solução.

Etapa 1: Revisão Bibliográfica e Estudo de Ferramentas

A primeira etapa consistiu em uma revisão abrangente da literatura relacionada ao monitoramento de processos industriais, *Big Data*, aprendizado de máquina e detecção de anomalias em poços de petróleo. Foram estudados artigos científicos,

dissertações e documentação técnica relevante, com foco especial nos seguintes materiais:

1. Artigos de referência sobre:
    a. Monitoramento baseado em desvios.
    b. Monitoramento e detecção de falhas baseado em dados.
    c. Projetos 3W e BibMon da Petrobras.
    d. Aplicação de algoritmos de aprendizado de máquina em processos industriais.
2. Documentação oficial da biblioteca BibMon, disponível em https://bibmon.readthedocs.io/.
3. Repositórios no GitHub:
    a. BibMon: https://github.com/petrobras/BibMon.
    b. Projeto 3W: https://github.com/petrobras/3W.
    c. Orientações para contribuição disponíveis na conta oficial da Petrobras.
4. Artigos de referência sobre metodologias *Big Data* e a arquitetura NIST.
5. Estudos sobre modelos de aprendizado de máquina usados para predição e classificação em séries temporais e documentação de bibliotecas relacionadas.

Essa etapa é fundamental para compreender os conceitos teóricos, as ferramentas disponíveis e as melhores práticas para o desenvolvimento da solução proposta.

Etapa 2: Planejamento do Desenvolvimento Utilizando Metodologias Ágeis

Com base no conhecimento adquirido, foi elaborado um plano de desenvolvimento seguindo os princípios de metodologias ágeis, adaptando o *framework Scrum* para o desenvolvimento individual. As atividades foram organizadas em sprints curtos, com objetivos claros e entregas contínuas, favorecendo a adaptação às eventuais mudanças de escopo e a incorporação de melhorias ao longo do processo.

A seguir, o plano de desenvolvimento do projeto, estruturado em sprints:

Sprints	**Duração**	**Objetivos**	**Atividades Principais**	**Entregáveis**
Sprint 1	2 semanas	Revisão Bibliográfica Estudo de Ferramentas	- Pesquisa de artigos científicos sobre monitoramento de processos industriais, *Big Data*, aprendizado de máquina e detecção de anomalias em poços de petróleo nos repositórios *ScienceDirect*, *Scopus* (*Elsevier*) e *IEEE*. - Estudo da documentação oficial da biblioteca BibMon e dos repositórios relacionados. - Análise de metodologias de *Big Data* e da arquitetura NIST.	- Relatório de revisão bibliográfica. - Lista de ferramentas e tecnologias relevantes para o projeto.
Sprint 2	1 semana	Planejamento do Projeto	- Definição do escopo teórico do projeto. - Organização das atividades de redação da monografia. - Elaboração do cronograma detalhado para a escrita.	- Plano de desenvolvimento da monografia. - Cronograma das atividades de redação.
Sprint 3	3 semanas	Redação da Fundamentação Teórica	- Leitura dos artigos selecionados e escrita da fundamentação teórica. - Discussão sobre a integração do *dataset* 3W e da biblioteca BibMon.	- Capítulo sobre fundamentação teórica da monografia redigido.
Sprint 4	4 semanas	Estruturação da instância de Big Data	- Detalhamento da proposta de instância de *Big Data* para monitoramento e detecção preditiva de anomalias. - Explicação dos métodos de E*nsemble* para integração de múltiplos modelos de aprendizado de máquina.	- Capítulo descrevendo a instância projetada.
Sprint 5	1 semana	Redação da Metodologia	- Descrição detalhada da metodologia adotada para o desenvolvimento do projeto teórico. - Explicação das etapas de pesquisa e planejamento.	- Capítulo de metodologia concluído.
Sprint 6	2 semanas	Análise de Resultados Esperados e Discussão	- Discussão sobre os resultados esperados com a implementação teórica da solução proposta. - Análise das contribuições potenciais e limitações do projeto.	- Capítulo de análise e discussão redigido.

Sprint 7	1 semana	Conclusão e Considerações Finais	- Síntese dos principais pontos abordados na monografia. - Apresentação das considerações finais e sugestões para trabalhos futuros.	- Capítulo de conclusão finalizado.
Sprint 8	2 semanas	Revisão e Formatação Final	- Revisão ortográfica e gramatical de todo o documento. - Adequação às normas acadêmicas e formatação conforme exigências institucionais.	- Monografia revisada e formatada, pronta para submissão.

Tabela 1 – Plano de desenvolvimento do projeto adaptado à *Sprints*

Etapa 3: Instanciação

A terceira etapa envolveu a instânciação do sistema *Big Data* para monitoramento e detecção de anomalias. Nesta fase, foram realizadas as seguintes atividades:

1. Adaptação da arquitetura NIST definida na etapa anterior, utilizando tecnologias apropriadas para processamento e armazenamento de grandes volumes de dados, como Apache Spark *(Apache Software Foundation, Wilmington, DE, EUA)*.
2. Integração do *dataset* 3W e da biblioteca BibMon à solução desenvolvida, adaptando-os conforme necessário para atender aos requisitos específicos do projeto.
3. Sugestão de uso de modelos de aprendizado de máquina para detecção de anomalias em séries temporais.
4. Estruturação de *pipelines* de dados para automatizar o fluxo desde a coleta até a análise e visualização, garantindo a eficiência e escalabilidade do sistema.

1.5 Organização do Trabalho

Esta monografia está organizada em quatro capítulos, cada um abordando aspectos específicos da pesquisa e desenvolvimento da solução.

No Capítulo 1, apresentamos a introdução ao tema, contextualizando a importância do monitoramento e detecção de anomalias em poços de petróleo e como as

tecnologias de *Big Data* e aprendizado de máquina podem aprimorar os processos atuais. Definimos o objetivo geral e os objetivos específicos da pesquisa, justificamos a relevância do trabalho e descrevemos a metodologia utilizada.

O Capítulo 2 aborda os principais conceitos e a fundamentação teórica que sustentam este projeto. Apresentamos os processos atuais de identificação de anomalias em poços de petróleo, incluindo técnicas de monitoramento baseado em desvios. Em seguida, incluímos o *dataset* 3W da Petrobras, a biblioteca BibMon e a arquitetura NIST para *Big Data*, contextualizando como esses elementos se relacionam com a solução proposta. Também detalhamos técnicas de *Ensemble*, as tecnologias de *Big Data* e os algoritmos de aprendizado de máquina para séries temporais que serão utilizados.

No Capítulo 3, discorremos sobre a instância de Big Data apresentada para monitoramento e detecção preditiva de anomalias em poços de petróleo. Apresentamos o diagrama da instância. Descrevemos as camadas de ingestão, armazenamento, processamento, análise e previsão de dados e finalizamos com a camada de visualização.

Por fim, no Capítulo 4, apontamos as conclusões da pesquisa, interpretando os resultados obtidos e discutindo as contribuições do trabalho para o avanço do conhecimento e da prática na área. Abordamos também as limitações da pesquisa e sugerimos possibilidades para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste trabalho aborda os conceitos e tecnologias essenciais para a construção de uma solução robusta para o monitoramento e a detecção preditiva de anomalias em poços de petróleo.

Inicialmente, apresentamos o contexto da indústria de óleo e gás e a importância do monitoramento de processos industriais, com foco na metodologia de detecção de desvios. Em seguida, aprofundamos a análise do *dataset* 3W da Petrobras, que servirá como base para a monitoramento de eventos indesejados e alertas em poços de petróleo reais. Posteriormente será apresentada a biblioteca BibMon da Petrobras, uma ferramenta de código aberto integrada às camadas de processamento e predição do sistema.

A arquitetura NIST para *Big Data* será indicada como modelo de referência para o desenvolvimento do sistema, juntamente com as tecnologias de *Big Data* que o sustenta, como Apache Airflow *(Apache Software Foundation, Wilmington, DE, EUA)*, para orquestração do fluxo de dados ou *Spark*, para processamento.

O aprendizado de máquina para séries temporais será apresentado como técnica central para a detecção de anomalias, com foco em algoritmos de classificação e técnicas de *Ensemble*. Por fim, serão apresentadas as métricas de avaliação para análise de desempenho da solução, garantindo a robustez e a confiabilidade do sistema.

2.1 Processos Atuais de Identificação de Falhas em Poços de Petróleo

Atualmente, diversas técnicas são empregadas para monitorar e detectar anomalias em poços de produção de petróleo, incluindo métodos estatísticos, modelos baseados em conhecimento e algoritmos de aprendizado de máquina. (MELO et al., 2024).

Tradicionalmente, a identificação de falhas se baseia na experiência de engenheiros e operadores, que analisam dados de sensores e indicadores de performance para detectar anomalias. Esses dados podem incluir:

1. Pressão: Pressão no anular, pressão a jusante e a montante dos *chokes*, pressão no fundo do poço (medida por sensores como o PDG - Permanent Downhole Gauge). (VARGAS et al., 2019)
2. Temperatura: Temperatura a jusante e a montante dos *chokes*, temperatura no fundo do poço. (VARGAS et al., 2019)
3. Vazão: Vazão de óleo, gás e água. (VARGAS et al., 2019)
4. Estado das válvulas: Posição (aberta/fechada) das válvulas de segurança, *chokes* e outras válvulas. (VARGAS et al., 2019)

A análise manual desses dados pode ser desafiadora, especialmente em poços complexos com grande volume de dados. O monitoramento multivariado, aliado a técnicas de *Machine Learning*, permitem analisar as interações entre diferentes variáveis e identificar anomalias que podem indicar a ocorrência de falhas. (MELO et al., 2024)

Falhas estas, que podem ser categorizadas em diversos tipos, como:

1. Falhas mecânicas: Problemas em equipamentos, como bombas, válvulas e tubulações. (MELO et al., 2024)
2. Problemas de fluxo: Alterações no fluxo de produção, como *slugging* e instabilidade de fluxo. (VARGAS et al., 2019)
3. Anomalias no reservatório: Mudanças nas características do reservatório, como aumento da produção de água ou perda de produtividade. (VARGAS et al., 2019)

Figura 1 - Hidrato em um poço da plataforma P-34 da Petrobras. Andreolli (2016)

Os dados dos sensores podem ser utilizados para identificar essas falhas. Por exemplo, um aumento repentino na temperatura pode indicar uma falha mecânica, enquanto uma queda na pressão pode indicar um problema de fluxo. A detecção precoce desses eventos é essencial para evitar perdas de produção, custos elevados de reparo e, em casos mais graves, acidentes com vítimas e impactos ambientais significativos.

A contribuição deste projeto aos processos atuais, dá-se pela integração do *dataset* 3W e da biblioteca BibMon a uma instância de Big Data que comporta técnicas de Ensemble para classificação de dados, aplicando simultaneamente, múltiplos modelos de *Machine Learning*. Ao fazê-lo, buscamos facilitar a disseminação do uso desses recursos, proporcionando à comunidade uma plataforma estruturada que simplifica o desenvolvimento e a validação de novos modelos de aprendizado de máquina. Essa integração não apenas agiliza o processo para profissionais e pesquisadores interessados em explorar os dados e as ferramentas, mas também incentiva a realização de trabalhos futuros e promove um ambiente mais propício para a inovação aberta. Espera-se que essa abordagem colaborativa beneficie não só a indústria de petróleo e gás, mas também outros setores que enfrentam desafios semelhantes relacionados ao volume e à complexidade de processos industriais de monitoramento e detecção de falhas.

2.1.1 Monitoramento de processos baseado em desvios

O monitoramento de processos baseado em desvios (*Deviation-Based Process Monitoring*) é uma técnica que visa capturar desvios em relação a valores ou padrões esperados em processos industriais. (MELO et al., 2024) Essa metodologia tem se tornado cada vez mais relevante na Indústria 4.0, à medida que a digitalização e a automação geram grandes volumes de dados que exigem uma análise eficiente e escalável.

Melo et al. (2024) descrevem a técnica como uma abordagem que compara medições reais de sensores com representações de modelos que representam o

comportamento normal do sistema. Desvios significativos entre os dados e o modelo podem indicar a ocorrência de anomalias.

Para construir o modelo de comportamento normal, podem ser utilizadas diferentes técnicas, como:

1. Modelos estatísticos: Médias móveis, análise de regressão, entre outros.
2. Modelos baseados em conhecimento: Regras e heurísticas definidas por especialistas.
3. Modelos de *Machine Learning:* Redes neurais, árvores de decisão, entre outros.

A comparação entre os dados e o modelo pode ser feita com base em diferentes métricas, como:

1. Erro quadrático de predição (SPE): Mede a diferença entre os valores reais e os valores previstos pelo modelo. (MELO et al., 2024)
2. Distância de *Mahalanobis*: Mede a distância multivariada entre um ponto de dados e a distribuição normal do modelo.
3. Taxa de falsos positivos e falsos negativos: Mede a taxa de erros na detecção de anomalias.

O monitoramento baseado em desvios permite a detecção precoce de anomalias, o que possibilita a tomada de medidas preventivas para evitar falhas e garantir a segurança e a eficiência do processo. (MELO et al., 2024)

2.2 O Dataset 3W

O *dataset* 3W é um conjunto de dados realista, disponibilizado pela Petrobras. Foi concebido em 2019 e lançado publicamente em 30 de maio de 2022 como parte do projeto 3W, sob a licença *Creative Commons* CC BY Atribuição 4.0 Internacional no repositório público da instituição no GitHub. O projeto 3W, parte de uma ação estratégica da companhia, liderada por seu departamento responsável pela Garantia de Fluxo e CENPES (Centro de Pesquisas, Desenvolvimento e Inovação Leopoldo

Américo Miguez de Mello), sendo o piloto de um programa denominado Conexões para Inovação - Módulo *Open Lab*, que busca soluções para os desafios de negócio da empresa, por meio da colaboração e da inovação aberta.

O *dataset* 3W contém informações reais geradas por sensores de monitoramento em poços de petróleo, dados simulados e dados desenhados à mão por especialistas. O conjunto de dados é composto por instâncias de nove diferentes tipos de eventos raros e indesejáveis, além de dados de operação normal dos poços, ambos caracterizados por múltiplas variáveis de processo.

Vargas et al. (2019) descrevem o *dataset* como um conjunto de dados desafiador que pode ser usado como referência para o desenvolvimento de técnicas de aprendizado de máquina e métodos para detecção e diagnóstico de eventos indesejáveis em poços de petróleo.

A instância projetada irá integrar somente os dados reais do *dataset* 3W fornecendo informações sobre 40 poços de petróleo, com variáveis de monitoramento coletadas no período entre 2011 e 2023, a partir de períodos de medições auferidas a cada segundo em diferentes janelas de tempo. Os eventos descritos compreendem dados de operações normais do poço, dados em estado transitórios e dados de falhas constatadas sendo; aumento abrupto de água produzida, fechamento espúrio da válvula de segurança de fundo de poço (DHSV), oscilações na produção de óleo e gás, instabilidade no fluxo de produção, perda rápida de produtividade do poço, restrição rápida no *choke* de produção (PCK), formação de incrustações no *choke* de produção e formação de hidratos na linha de produção e na linha de serviço.

2.3 A Biblioteca BibMon

A BibMon (Biblioteca de Monitoramento de Processos) é uma biblioteca Python de código aberto desenvolvida para auxiliar na tarefa de monitoramento de processos, detecção de falhas e desenvolvimento de sensores virtuais. A biblioteca originou-se de projetos de pesquisa conduzidos em colaboração entre o Programa de Engenharia Química da COPPE/UFRJ e o CENPES/Petrobras e oferece diversas ferramentas e funcionalidades para análise de dados, incluindo modelos de regressão e

reconstrução, pipelines de pré-processamento, alarmes e visualização de dados. (MELO et al., 2023).

De acordo com Melo et al. (2023), a BibMon foi projetada para ser extensível e de fácil manutenção, permitindo a integração de novos modelos e metodologias. A biblioteca oferece modelos como PCA (Análise de Componentes Principais), ESN (Redes Neurais de Estado de Eco) e algoritmos baseados em similaridade, além de recursos para pré-processamento de dados e geração de gráficos de controle.

A instância proposta integra a biblioteca BibMon à *pipeline* de dados, nas camadas de processamento e predição, para auxiliar no monitoramento e detecção de falhas em poços de petróleo. A BibMon será utilizada para:

1. Pré-processamento dos dados: Limpeza, tratamento de valores ausentes e normalização dos dados, utilizando as funções da classe PreProcess. (MELO et al., 2023)
2. Criação e treinamento de modelos: Implementação e treinamento de modelos de *Machine Learning*, como Random Forest e Regressão Linear, utilizando a classe sklearnRegressor. (MELO et al., 2023)
3. Avaliação do desempenho dos modelos: Cálculo de métricas de desempenho, como SPE (*Squared Prediction Error*) e FDR (*Fault Detection Rate*), utilizando as funções plot_SPE e plot_predictions. (MELO et al., 2023)
4. Geração de gráficos e visualizações: Criação de gráficos de controle e mapas de diagnóstico para auxiliar na interpretação dos resultados e na identificação de anomalias, utilizando as funções plot_SPE, plot_predictions e spearmanr_dendrogram. (MELO et al., 2023)

A integração da BibMon à instância de Big Data permitirá o desenvolvimento de um sistema mais robusto e eficiente para o monitoramento e a detecção de falhas em poços de petróleo, aproveitando as funcionalidades da biblioteca para pré-processamento, modelagem, avaliação e visualização dos dados.

2.4 Arquitetura NIST de referência para Big Data

O Instituto Nacional de Padrões e Tecnologia (NIST), dos Estados Unidos, busca impulsionar a inovação e a competitividade industrial por meio de colaborações internacionais para estabelecer padrões de referência e conduzir medições precisas. (NIST, 2019)

Em 2013, o NIST formou o Grupo de Trabalho Público de *Big Data* (NBD-PWG) com o objetivo de promover o avanço do *Big Data*. O NBD-PWG, composto por representantes da indústria, universidades e governo, documentou suas atividades em uma série de sete volumes, que abordam tópicos como definições, taxonomia, segurança, privacidade e arquitetura de *Big Data*. (NIST, 2019)

O NBD-PWG propôs um modelo conceitual de arquitetura de referência para *Big Data*, independente de fornecedor, tecnologia e infraestrutura. Esse modelo, conhecido como NBDRA (*NIST Big Data Reference Architecture*), descreve as características, os sistemas e a análise de *Big Data*, considerando a relação custo-benefício. (NIST, 2019)

A NBDRA é um guia para a construção de sistemas de *Big Data*, definindo os seguintes componentes e suas interações:

1. Sistema Orquestrador: Define os requisitos e políticas para o sistema de *Big Data*, monitorando a conformidade e o ciclo de vida dos dados. (NIST, 2019)
2. Provedor de Dados: Introduz dados no sistema, provendo acesso para processamento e análise. (NIST, 2019)
3. Provedor de Aplicações de Big Data: Executa o processamento e a análise, transformando e extraindo informações. (NIST, 2019)
4. Provedor de *Framework* de *Big Data*: Fornece a infraestrutura para o sistema, incluindo plataformas, armazenamento e gerenciamento. (NIST, 2019)
5. Consumidor de Dados: Utiliza os resultados para gerar *insights*, tomar decisões e criar valor. (NIST, 2019)

A NBDRA também define duas camadas que interagem com todos os componentes:

1. Camada de Gerenciamento: Responsável por gerenciar os recursos do sistema, incluindo provisionamento, configuração, monitoramento e gerenciamento do ciclo de vida dos dados. (NIST, 2019)
2. Camada de Segurança e Privacidade: Responsável por proteger e garantir a privacidade dos dados, incluindo autenticação, autorização e auditoria. (NIST, 2019)

A Figura 1 ilustra a arquitetura de referência NIST para *Big Data*, com seus componentes e camadas.

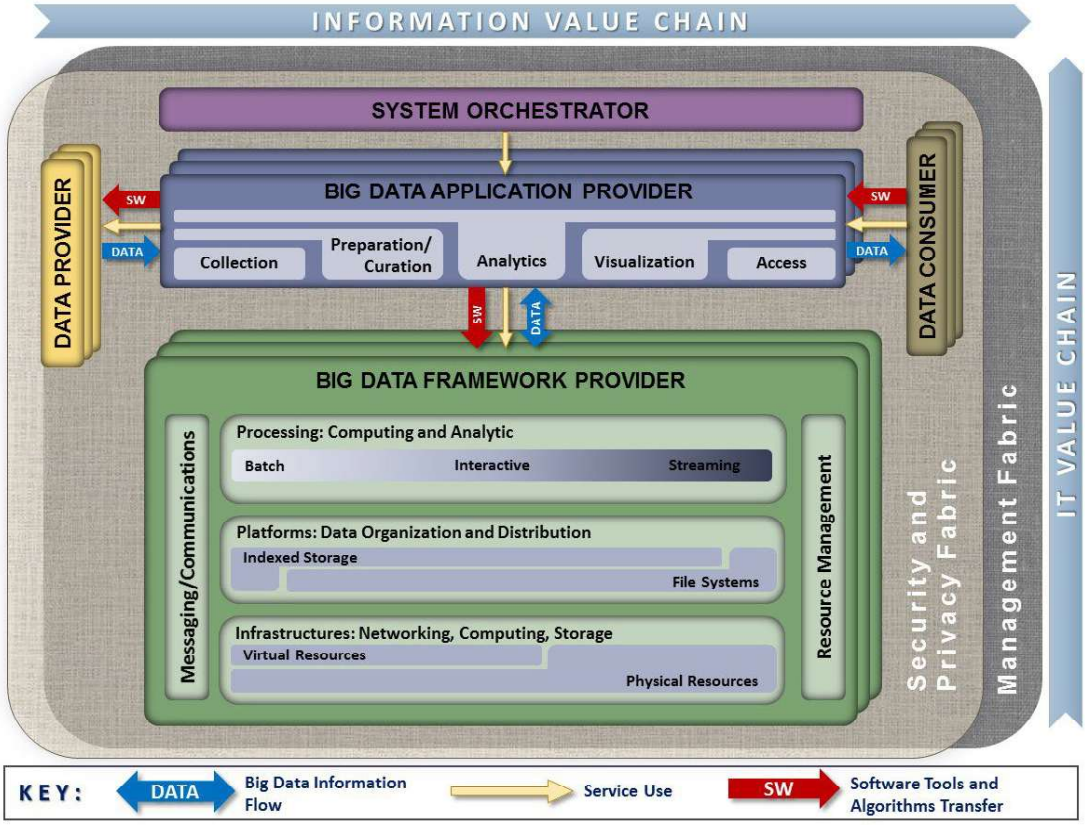

Figura 2 - Arquitetura de referência NIST para *Big Data* (NIST, 2019)

2.5 Tecnologias de Big Data

As tecnologias de *Big Data* são ferramentas e plataformas projetadas para lidar com grandes volumes de dados, alta velocidade de ingestão e variedade dos dados. Essas

tecnologias permitem a coleta, o armazenamento, o processamento e a análise de dados em escala.

A solução *Big Data* proposta para detecção de anomalias em poços de petróleo utiliza diversas tecnologias, que podem ser categorizadas de acordo com os papéis definidos na Arquitetura de Referência NIST para *Big Data* (NBDRA):

1. Sistema Orquestrador:

No projeto em questão, o Apache Airflow, em sua versão 2.10, será incorporado ao provedor de *framework* através do Amazon Managed Workflows for Apache Airflow (MWAA), um serviço gerenciado da AWS que facilita a execução de *workflows* do Apache Airflow em nuvem. O MWAA oferece uma experiência "*Airflow as a Service*", simplificando a configuração, a operação e a escalabilidade do Airflow. (Amazon MWAA Documentation, 2024)

O Apache Airflow é uma plataforma de código aberto para orquestrar *workflows* de forma programática. Ele permite definir, agendar e monitorar fluxos de trabalho complexos, como *pipelines* de dados, de forma visual e intuitiva. (Apache Airflow Documentation, 2024)

As tarefas do *workflow* são definidas como DAGs (*Directed Acyclic Graphs*), que são grafos acíclicos direcionados que representam a sequência de operações a serem executadas. O Airflow agenda e monitora a execução das DAGs, garantindo que as tarefas sejam executadas na ordem correta e no momento certo.

A utilização do Airflow como sistema orquestrador garante a automação, a organização e a confiabilidade de lidar com o grande volume de dados do *dataset* 3W. Permite o versionamento das etapas orquestradas, o desenvolvimento colaborativo e o monitoramento da execução das tarefas, gerando alertas em caso de falhas.

2. Provedor de Dados:

A solução utilizará o *Amazon Simple Storage Service* (*Amazon S3*). Um serviço de armazenamento de objetos que oferece escalabilidade, disponibilidade de dados, segurança e performance líderes do setor. (Amazon, 2024)

No Amazon S3, os dados são armazenados em *buckets*, que são containers para objetos. Os arquivos Parquet que contêm os dados dos poços de petróleo podem ser

diretamente mapeados para objetos no S3. O Amazon S3 atua então, como o repositório central para todos os dados do projeto, incluindo dados brutos, processados e resultados de predições.

Para Cleveland (2023), o Amazon S3 foi criado para armazenar e recuperar qualquer quantidade de dados a qualquer momento, de qualquer lugar. Sua durabilidade e disponibilidade o tornam uma escolha atraente para cargas de trabalho de *Big Data*. Além disso, ele oferece uma ampla gama de integrações com estruturas de *Big Data*.

3. Provedor de Aplicações de *Big Data*:

Esse componente é responsável por executar o processamento e a análise dos dados. Na solução proposta, essa função é realizada por uma combinação de tecnologias, incluindo o Apache Spark, para processamento distribuído e análise de dados em larga escala e a biblioteca BibMon, para pré-processamento de dados, implementação de algoritmos de *Machine Learning* e geração de visualizações.

O Apache Spark é um sistema de processamento distribuído de código aberto usado para *workloads* de *Big Data*. Ele será utilizado em várias camadas da solução, desde a conectividade rápida para acessar os dados brutos no Amazon S3, até a integração com o Catálogo de Dados do AWS Glue (serviço de integração de dados com tecnologia sem servidor que facilita a descoberta, preparação, movimentação e integração de dados de várias fontes para análise, *Machine Learning* e desenvolvimento de aplicações). (Amazon, 2024)

A integração do Spark e da BibMon garante que a solução seja capaz de processar grandes volumes de dados de forma eficiente e que os algoritmos de *Machine Learning* sejam aplicados de forma otimizada.

O Amazon EMR (anteriormente chamado de Amazon Elastic MapReduce) é uma plataforma de cluster gerenciada que simplifica a execução de estruturas de *Big Data*, como o Apache Spark, na AWS para processar e analisar grandes quantidades de dados. (Amazon, 2024) Será utilizado para transformar e mover grandes quantidades de dados para dentro e para fora da camada de armazenamento com Amazon Simple Storage Service (Amazon S3).

Amazon SageMaker Spark é uma biblioteca Spark de código aberto que irá criar *pipelines* de aprendizado de máquina do Spark. Isso simplificará a integração dos estágios, como treinamento e hospedagem de modelos.

4. Provedor de *Framework* de *Big Data*:

O Provedor de *Framework* de *Big Data* é a infraestrutura que suporta o sistema. A instância apresentada utiliza a plataforma Amazon Web Services (AWS) como plataforma de computação em nuvem, incluindo:

- Processamento e Análise de Dados:
  a) Apache Spark no Amazon EMR: Plataforma de processamento distribuído para *Big Data*, utilizada para processamento em cluster.
  b) AWS Glue: Serviço de integração de dados *serverless* que facilita a preparação e movimentação de dados.
  c) Amazon SageMaker: Serviço para construir, treinar e implantar modelos de *Machine Learning* em escala.
- Armazenamento de Dados:
  a) Amazon S3 (Simple Storage Service): Serviço de armazenamento de objetos escalável e durável.
- Organização e Distribuição de Dados:
  a) AWS Glue Data Catalog: Catálogo centralizado de metadados para dados armazenados no S3.
  b) Amazon Athena: Serviço de consulta interativa que permite executar SQL diretamente no S3 e oferece suporte ao formato Parquet.
  c) Amazon SageMaker Canvas: Interface visual sem código que permite preparar dados, criar e implantar modelos de *Machine Learning* altamente precisos, utilizando métodos *Ensemble*, simplificando o ciclo de vida de *Machine Learning* de ponta a ponta.
- Infraestrutura de Rede e Computação:
  a) Amazon VPC (Virtual Private Cloud): Provisão de uma rede virtual isolada para os recursos AWS.
  b) Auto Scaling Groups: Para ajustar automaticamente a capacidade computacional conforme a demanda.
- Mensageria e Comunicação:

a) DAG (*Directed Acyclic Graph*): Conceito centro do Airflow, reúne tarefas, organizadas com dependências e relacionamentos para dizer como elas devem ser executadas.
b) SageMakerEndpointOperator: Recurso do Airflow para enviar dados processados aos *endpoints* (Ponto *URL* de entrada para o serviço AWS).

- Gerenciamento de Recursos:
  a) AWS Config: Garante conformidade com políticas internas e regulamentações.
  b) AWS IAM (Identity and Access Management): Gerenciamento de acesso e permissões.
  c) AWS CloudWatch: Monitoramento e coleta de métricas dos recursos e aplicações.

A utilização da AWS como Provedor de *Framework* garante a escalabilidade, a disponibilidade e a segurança do sistema.

5. Consumidor de Dados:

O Consumidor de Dados é o usuário final do sistema, que pode ser um engenheiro, um operador, um gerente da área de produção de petróleo ou a comunidade de cientistas e analistas de dados.

A solução proposta disponibiliza os resultados do processamento e da análise de dados por meio de pontos que permitem ao consumidor de dados:

- Realizar consultas SQL para extrair subconjuntos específicos dos dados conforme necessário.
- Popular aplicações *Upstream*.
- Modelos treinados são implantados e podem consumir novos dados para predições, acessando novamente o S3 ou recebendo dados em tempo real.
- Conectar-se a soluções de visualizações de dados e disparo de alertas.

Ao seguir a arquitetura de referência NIST, a solução *Big Data* proposta para detecção e monitoramento de falhas em poços de petróleo garante a organização, a estruturação e a escalabilidade do sistema, além de facilitar a análise de requisitos e o desenvolvimento de uma solução eficiente e robusta.

2.6 *Machine Learning* e a abordagem *Ensemble*

O aprendizado de máquina para séries temporais se mostra promissor para a detecção de anomalias em processos industriais, especialmente em poços de petróleo, onde os dados são coletados sequencialmente ao longo do tempo. Algoritmos de *Machine Learning* podem ser utilizados para analisar e modelar dados de séries temporais, como pressão, temperatura e vazão, para identificar padrões e desvios que indicam anomalias. No contexto deste trabalho, o objetivo é utilizar modelos combinados de *Machine Learning* na camada de predição, Método *Ensemble*, para detectar anomalias que podem indicar falhas nos poços. Cha Zhang, 2012 apresenta como os métodos *Ensemble* são eficientes e versáteis para solucionar uma grande variedade de problemas relacionados à classificação, além de reduzir a variância e aumentar a acurácia de sistemas de tomada de decisão.

2.6.1 Análise do Dataset 3W no Contexto de Séries Temporais

O *dataset* 3W é uma série temporal multivariada, pois cada observação no tempo inclui múltiplas variáveis medidas simultaneamente. (MORETTIN; et al., 2023) Essas variáveis incluem pressão, temperatura, vazão e estado das válvulas, que juntas caracterizam o comportamento dinâmico de um poço de petróleo. A natureza multivariada do *dataset* é crucial para capturar as complexas interações entre as diferentes variáveis e, assim, detectar anomalias que podem ser indicativas de falhas.

As séries temporais do 3W são, em geral, não lineares e discretas. Essa não linearidade é esperada em sistemas complexos como poços de petróleo, onde diversos fatores físicos, químicos e geológicos interagem de maneira complexa. Apesar de o tempo ser contínuo, as medições são registradas em momentos específicos, formando uma sequência discreta de dados. Em algumas séries temporais do 3W, é possível observar sazonalidade. No caso dos poços de petróleo, a sazonalidade pode estar relacionada aos ciclos de produção, de injeção de gás ou água, ou a outros fatores que influenciam a produção de forma periódica.

Principalmente, observa-se que as séries temporais multivariadas, não lineares, de atributos discretos e contínuos do dataset 3W estão sujeitas a ruídos, *outliers* e

mudanças abruptas que podem ser indicativos de erros de medição, eventos externos, como intervenções na operação do poço, ou indicar anomalias que podem gerar falhas, e encontrar estes padrões é justamente o objetivo do uso, na modelagem preditiva, de algoritmos de aprendizado de máquina ou de aprendizado profundo.

Parte importante dessa instância de *Big Data* é ser capaz de usar diferentes formas para analisar correlações positivas, negativas ou nulas na modelagem de dados, com o objetivo de entender o comportamento do sistema e selecionar as *features* mais relevantes evitando redundâncias e multicolinearidade. A biblioteca BibMon oferece ferramentas para auxiliar na análise de correlação, como a função spearmanr_dendrogram, que gera um dendrograma de correlações de *Spearman*. (MELO et al., 2023, p. 4)

2.6.2 Algoritmos suportados pela solução

Algumas técnicas que podem ser consideradas, e que são suportadas pela solução, são:

- Modelos de classificação: Algoritmos como Árvores de Decisão ou mesmo o modelo *Random Forest* podem ser utilizados para classificar as observações como normais ou anômalas. O algoritmo *Random Forest* é geralmente robusto a alta dimensionalidade, pois ele cria múltiplas árvores de decisão que consideram diferentes subconjuntos de features. Isso ajuda a reduzir o *overfitting* e a capturar as interações entre as variáveis.

- *Deep Learning*: Redes Neurais Recorrentes (RNN), como LSTM (*Long Short Term Memory Unit*), podem ser eficazes na detecção de anomalias em séries temporais, pois conseguem capturar relações não lineares complexas nos dados. As RNNs podem ser especialmente úteis para lidar com a não linearidade e a alta frequência de amostragem do *dataset* 3W.

A seleção do algoritmo de *Machine Learning* mais adequado para cada tipo de evento anômalo deve ser feita com base em uma análise criteriosa do *dataset* 3W e dos

requisitos do sistema. A alta dimensionalidade do *dataset* 3W pode ser um desafio para qualquer modelo utilizado. Técnicas de redução de dimensionalidade, como PCA (Análise de Componentes Principais), podem ser aplicadas aos dados pré-processados para gerar os componentes principais, reduzir o número de *features* e melhorar o desempenho dos modelos. A BibMon facilita a aplicação do PCA e a integração com os modelos de *Machine Learning* através da classe PCA para criar um modelo PCA. O método fit pode ser usado para treinar o modelo com os dados pré-processados e, o atributo components_, para obter os componentes principais e transformar os dados originais nos novos componentes principais selecionados com o método transform. Assim, os dados transformados (com dimensionalidade reduzida) podem ser utilizados para treinar os modelos de *Machine Learning* selecionados.

A BibMon oferece funções para visualizar os resultados do PCA, como plot_cumulative_variance e plot_SPE, que ajudam na análise e interpretação dos dados.

2.6.3 Abordagem Ensemble

O uso de métodos *Ensemble* em *Machine Learning* é amplamente reconhecido por sua capacidade de melhorar a precisão e a robustez de modelos preditivos. No contexto do *dataset* 3W, que inclui dados complexos e multivariados para detecção de falhas em poços de petróleo, o *Ensemble* torna-se especialmente relevante, pois permite combinar as vantagens de múltiplos algoritmos para mitigar erros associados a variância e viés.

Na figura 2, apresentamos um “funil” conceitual que ilustra a jornada dos dados desde a produção por sensores, em poços de petróleo, até a geração de *insights* estratégicos. Nele, as informações coletadas em campo passam por uma plataforma robusta de *Big Data*, onde são agregadas e processadas em larga escala. Os dados transformados são enviados a uma camada de predição onde, múltiplos modelos de *Machine Learning* (várias redes neurais ou algoritmos especializados), organizados em um método *ensemble*, classificam o estado de operação e identificam possíveis anomalias. Finalmente, no nível superior, surgem os resultados ou *endpoints* de inferência, tais como previsões (*Predictions*), análise de causa-raiz (*Root Cause*

Analysis), análise de tendências (*Trend Analysis*), detecção de anomalias (*Anomaly Detection*) e *meta-learning*. Essa disposição em camadas lembra a trajetória dos dados: da coleta bruta até a entrega de *insights* avançados.

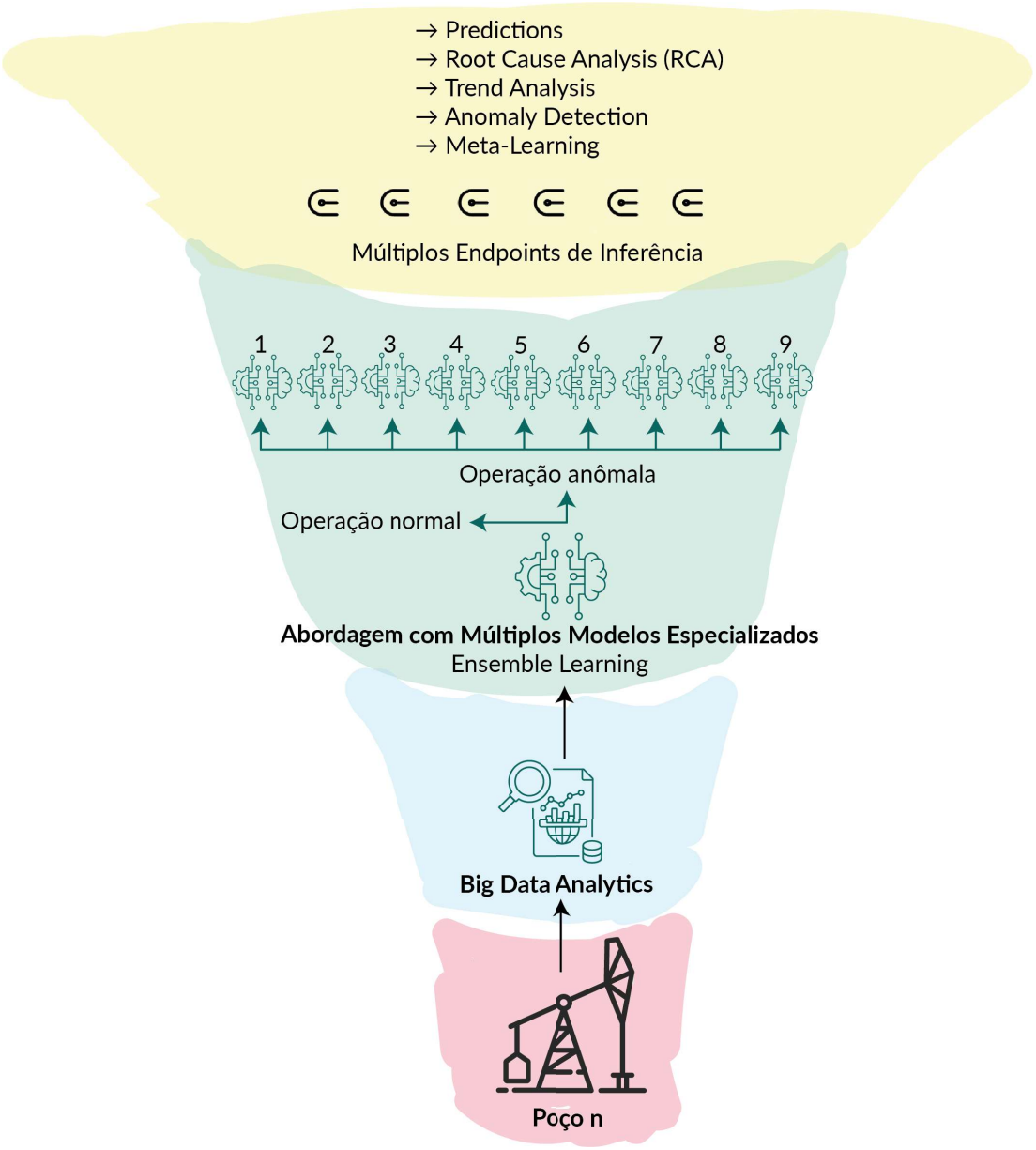

Figura 3 - Jornada dos dados desde a produção, por sensores instalados em poços de petróleo, até a geração de *insights* estratégicos sob abordagem *Ensemble*. (O autor, 2024).

Técnicas de *Ensemble*, como *Bagging*, *Boosting* e *Stacking*, oferecem soluções adaptáveis para problemas inerentes ao *dataset* 3W como, alta dimensionalidade, desbalanceamento de classes e correlação entre variáveis.

Bagging: Divide o *dataset* em subconjuntos para treinar múltiplos modelos independentes. No contexto do 3W, isso pode ser implementado no Amazon SageMaker, utilizando o algoritmo *Random Forest*. Cada modelo treina em diferentes subconjuntos dos dados, reduzindo a variância e melhorando a robustez geral.

Boosting: Ajusta o peso das observações mal classificadas em iterações sucessivas, permitindo que o modelo aprenda com erros anteriores. Técnicas como o XGBoost podem ser configuradas no SageMaker para explorar este método, particularmente útil para capturar padrões sutis em falhas raras.

Stacking: Utiliza múltiplos modelos base para gerar previsões que são então combinadas por um meta-modelo. Essa abordagem, configurada no SageMaker com *frameworks* como TensorFlow ou Scikit-learn, pode combinar a força de diversos algoritmos, criando um modelo final mais robusto e generalizável.

A abordagem *Ensemble* aplicada ao *dataset* 3W é projetada para maximizar a precisão na detecção de anomalias enquanto reduz falsos positivos e negativos. A combinação de algoritmos por meio de Stacking é particularmente promissora, dada sua flexibilidade e capacidade de incorporar diferentes perspectivas analíticas em um único modelo.

2.6.4 Métricas de avaliação

A avaliação da performance da solução de *Big Data* proposta para detecção de anomalias em poços de petróleo exige uma análise abrangente que considere diferentes aspectos, indo além da simples avaliação dos modelos de *Machine Learning*. As métricas de avaliação devem refletir a capacidade da solução de lidar com as características específicas do *dataset* 3W, os requisitos do sistema e os desafios inerentes ao processamento de grandes volumes de dados.

2.6.4.1 Métricas da Solução *Big Data*

Escalabilidade: A solução deve ser capaz de processar grandes volumes de dados de forma eficiente, acomodando o crescimento futuro do *dataset* e a necessidade de análises mais complexas. A escalabilidade pode ser medida em termos de volume de

dados processados por unidade de tempo, número de usuários simultâneos suportados e capacidade de resposta do sistema sob diferentes cargas de trabalho. (NIST, 2019)

Eficiência: O tempo de processamento e a utilização de recursos computacionais devem ser otimizados para garantir a detecção de anomalias em tempo real e evitar atrasos que podem prejudicar a tomada de decisão. A eficiência pode ser medida em termos de tempo de resposta do sistema, consumo de *CPU* e memória, e custo de processamento por unidade de dado. (NIST, 2019)

Robustez: A solução deve ser resiliente a falhas e robusta o suficiente para lidar com dados incompletos ou inconsistentes. A robustez pode ser medida em termos de tempo de recuperação em caso de falhas, capacidade de lidar com dados faltantes ou corrompidos e taxa de erros no processamento de dados. (NIST, 2019)

Manutenibilidade: O sistema e o código-fonte devem ser claros, organizados e bem documentados para facilitar a manutenção e a atualização da solução. A manutenibilidade pode ser medida em termos de tempo e esforço necessários para corrigir erros, implementar novas funcionalidades e atualizar o sistema.

Usabilidade: A interface de visualização e os dashboards devem ser intuitivos e fáceis de usar, permitindo que os usuários finais compreendam os resultados e tomem decisões de forma rápida e eficiente. A usabilidade pode ser medida por meio de testes com usuários, avaliando a facilidade de aprendizado, a eficiência na realização de tarefas e a satisfação dos usuários. (NIST, 2019)

Segurança: A solução deve garantir a segurança e a privacidade dos dados, implementando mecanismos de autenticação, autorização e criptografia para proteger as informações confidenciais. A segurança pode ser medida em termos de número de tentativas de acesso não autorizado, taxa de sucesso em ataques de segurança, tempo de recuperação em caso de incidentes de segurança e conformidade com a legislação de privacidade de dados local. (NIST, 2019)

Custo-benefício: A implementação e a operação da solução devem ser economicamente viáveis, considerando o custo da infraestrutura, dos serviços e da equipe de desenvolvimento e manutenção. O custo-benefício pode ser medido em

termos de retorno sobre o investimento, tempo de retorno do capital investido e economia de custos em relação aos métodos tradicionais de detecção de anomalias. (NIST, 2019)

2.6.4.2 Métricas dos Modelos de *Machine Learning*

As métricas de avaliação dos modelos de *Machine Learning* também são importantes para a avaliação da performance da solução *Big Data*, pois fornecem informações sobre a capacidade do sistema de detectar anomalias com precisão. Algumas métricas relevantes são:

Erro Quadrático de Predição (SPE): Mede a diferença entre os valores reais e os valores previstos pelo modelo, avaliando a qualidade da previsão. (MELO et al., 2023)

Taxa de Detecção de Falhas (FDR): Mede a proporção de anomalias corretamente identificadas pelo sistema, avaliando a sensibilidade do modelo na detecção de eventos anômalos. (MELO et al., 2023)

Acurácia: A acurácia será utilizada para avaliar a capacidade do modelo de classificar corretamente as amostras como normais ou anômalas. Uma alta acurácia indica que o modelo é capaz de distinguir entre os diferentes estados do poço.

2.6.4.3 Relação entre as Métricas

As métricas dos modelos de *Machine Learning*, como SPE, FDR e acurácia, fornecem informações importantes para a avaliação da performance da solução *Big Data* como um todo. Por exemplo, uma alta taxa de falsos positivos nos modelos pode indicar a necessidade de ajustes na etapa de pré-processamento dos dados ou na escolha dos algoritmos, impactando a robustez e a eficiência da solução.

A escolha das métricas de avaliação deve ser feita de forma criteriosa, considerando os objetivos do projeto, as características do *dataset* 3W e os requisitos da solução *Big Data*. A utilização de diferentes métricas, que avaliam tanto os modelos de *Machine Learning* quanto a solução como um todo, permitem uma análise mais completa e abrangente da performance do sistema.

3 INSTÂNCIA DE BIG DATA PARA O MONITORAMENTO DE ANOMALIAS E A DETECÇÃO PREDITIVA DE FALHAS EM POÇOS DE PETRÓLEO, UTILIZANDO APRENDIZAGEM DE MÁQUINA

Visando atender à necessidade da indústria de petróleo e gás de monitorar e detectar anomalias em poços de petróleo de forma eficiente e escalável, estruturamos a instância de *Big Data* em camadas, integrando diferentes tecnologias e ferramentas para coletar, armazenar, processar, analisar e visualizar dados. A disposição do sistema é baseada na Arquitetura de Referência NIST para *Big Data* (NBDRA), que fornece um modelo conceitual para a organização e estruturação de sistemas de *Big Data*. (NIST, 2019)

A solução integra o *dataset* 3W da Petrobras, que fornece dados reais de poços de petróleo, a biblioteca BibMon da Petrobras, que oferece ferramentas para pré-processamento, análise e modelagem de dados e tecnologias de *Big Data* provisionadas pelo serviço de computação em nuvem da AWS (Amazon Web Services).

3.1 Diagrama da Instância

A figura 2 a seguir, apresenta um diagrama com a visão geral, de alto nível, da instância de *Big Data* para monitoramento e detecção de falhas em poços de petróleo baseada na arquitetura de referência NIST para *Big Data*. (NIST, 2019)

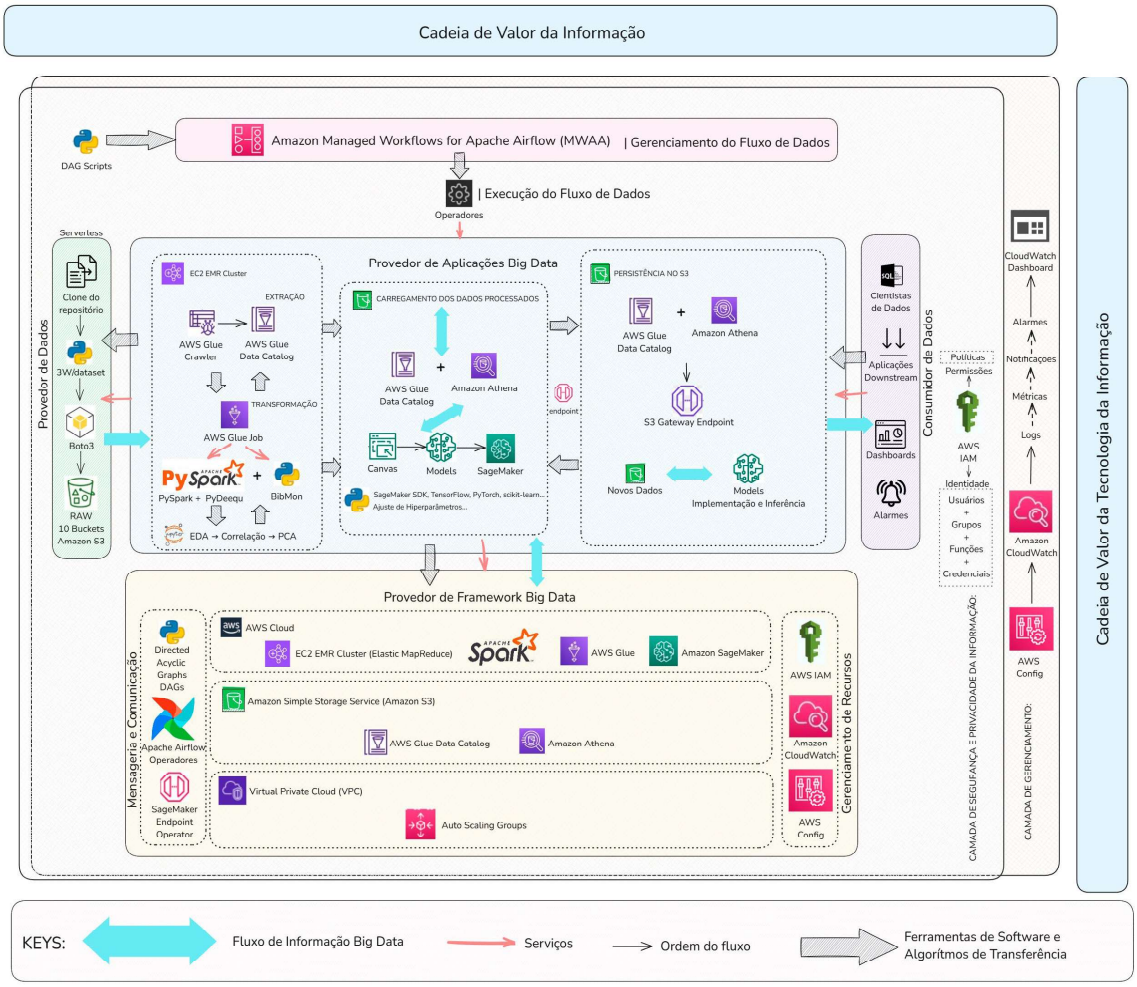

Figura 4 – Diagrama da instância de *Big Data* para monitoramento e detecção de falhas em poços de petróleo. O autor, 2024.

A seguir, um resumo do comportamento do sistema:

1. Ingestão e Catalogação:
   1.1. O AirFlow inicia o processo e aciona o AWS Glue Crawler.
   1.2. O Crawler escaneia os dados no S3 e atualiza o Data Catalog.
   1.3. O Dicionário de Dados é sincronizado com as novas informações.
2. Processamento e Transformação:
   2.1. Scripts PySpark são executados via AWS Glue Jobs, orquestrados pelo AirFlow.
   2.2. A BibMon é integrada para pré-processamento e análise.
   2.3. Dados são limpos, normalizados e transformados.

3. Modelagem e Predição:
   3.1. Modelos de Machine Learning são treinados utilizando BibMon e PySpark.
   3.2. Avaliação e validação dos modelos são realizadas.
   3.3. Modelos e resultados são armazenados no S3.
4. Carregamento e Disponibilização:
   4.1. Dados processados são escritos de volta no S3.
   4.2. O Data Catalog é atualizado para refletir as mudanças.
   4.3. Resultados são disponibilizados para visualização e consumo.
5. Visualização e Monitoramento:
   5.1. Disponibilizado *endpoint* para consulta SQL direta ou consumo por aplicações *downstream*.
   5.2. *Dashboards* interativos permitem a análise dos resultados.
   5.3. Alertas são configurados para notificações em tempo real.
   5.4. O sistema é monitorado e mantido para garantir desempenho e segurança.

3.2 Fontes de dados

Disponibilizado publicamente pelo repositório da Petrobras no GitHub (https://github.com/petrobras/3W/), o *dataset* 3W, em sua versão 2.0.0, lançado em 25 de julho de 2024, possui 1,74GB e contém 2228 arquivos Parquet distribuídos em 10 pastas, representando diferentes tipos de eventos e cenários.

O formato de armazenamento em colunas e de código aberto, características básicas dos arquivos Parquet, são ideais para cenários de *Big Data* e consultas analíticas. Um arquivo Parquet armazena colunas juntas para que os bancos de dados possam retornar informações de uma coluna específica mais rapidamente, em vez de pesquisar em cada linha com várias colunas.

O *dataset* 3W está dividido em 10 pastas numeradas sequencialmente de 0 a 9, representando cada uma, tipos específicos de anomalias em poços de petróleo:

Pasta	**Evento**	**Descrição**
0	NORMAL	Operação normal do poço.
1	ABRUPT_INCREASE_OF_BSW	Aumento abrupto de água produzida.
2	SPURIOUS_CLOSURE_OF_DHSV	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV).
3	SEVERE_SLUGGING	*Slug flow* severo (oscilações na produção de óleo e gás).
4	FLOW_INSTABILITY	Instabilidade no fluxo de produção.
5	RAPID_PRODUCTIVITY_LOSS	Perda rápida de produtividade do poço.
6	QUICK_RESTRICTION_IN_PCK	Restrição rápida no *choke* de produção (PCK).
7	SCALING_IN_PCK	Formação de incrustações no *choke* de produção.
8	HYDRATE_IN_PRODUCTION_LINE	Formação de hidratos na linha de produção.
9	HYDRATE_IN_SERVICE_LINE	Formação de hidratos na linha de serviço.

Tabela 2 – Pastas do *dataset* 3W separadas por tipo de evento anômalo.

Cada uma dessas pastas contém dezenas de arquivos e cada arquivo é capaz de armazenar informações observadas em diferentes intervalos de dias, auferidas a cada segundo, podendo conter milhares de linhas cada arquivo.

São 28 as variáveis de processo observadas e todos os poços, além da data da observação:

Variável	**Unidade**	**Descrição**
Timestamp	seg	Instante em que a observação foi gerada.
ABER-CKGL	%	Abertura do *choke* de gás *lift*.
ABER-CKP	%	Abertura do *choke* de produção.
ESTADO-DHSV	-	Estado da válvula de segurança de fundo de poço (valores possíveis: 0, 0.5 ou 1).
ESTADO-M1	-	Estado da válvula mestre de produção (valores possíveis: 0, 0.5 ou 1).
ESTADO-M2	-	Estado da válvula mestre do anular (valores possíveis: 0, 0.5 ou 1).
ESTADO-PXO	-	Estado da válvula de crossover de *pig* (valores possíveis: 0, 0.5 ou 1).

ESTADO-SDV-GL	-	Estado da válvula de shutdown de gás *lift* (valores possíveis: 0, 0.5 ou 1).
ESTADO-SDV-P	-	Estado da válvula de shutdown de produção (valores possíveis: 0, 0.5 ou 1).
ESTADO-W1	-	Estado da válvula de asa de produção (valores possíveis: 0, 0.5 ou 1).
ESTADO-W2	-	Estado da válvula de asa do anular (valores possíveis: 0, 0.5 ou 1).
ESTADO-XO	-	Estado da válvula de crossover (valores possíveis: 0, 0.5 ou 1).
P-ANULAR	Pa	Pressão no anular do poço.
P-JUS-BS	Pa	Pressão a jusante da bomba de serviço.
P-JUS-CKGL	Pa	Pressão a jusante do *choke* de gás *lift*.
P-JUS-CKP	Pa	Pressão a jusante do *choke* de produção.
P-MON-CKGL	Pa	Pressão a montante do *choke* de gás *lift*.
P-MON-CKP	Pa	Pressão a montante do *choke* de produção.
P-MON-SDV-P	Pa	Pressão a montante da SDV de produção.
P-PDG	Pa	Pressão no medidor permanente de fundo de poço.
PT-P	Pa	Pressão a jusante da válvula de asa de produção no tubo de produção.
P-TPT	Pa	Pressão no transdutor de temperatura e pressão.
QBS	m^3/s	Vazão na bomba de serviço.
QGL	m^3/s	Vazão de gás *lift*.
T-JUS-CKP	°C	Temperatura a jusante do *choke* de produção.
T-MON-CKP	°C	Temperatura a montante do *choke* de produção.
T-PDG	°C	Temperatura no medidor permanente de fundo de poço.
T-TPT	°C	Temperatura no transdutor de temperatura e pressão.

Tabela 3 – Variáveis monitoradas, unidade de medida e descrição.

Também dois rótulos de informações sobre os estados dos poços de petróleo:

class: Rótulo da observação, indicando a classe de evento.

a) 0 = Operação Normal
b) 1 a 9 = Estado de anomalia
c) 101 a 109 = Estado de transição para uma anomalia

state: Estado operacional do poço. Aberto ou fechado.

Os arquivos dentro das pastas são nomeados de acordo com a seguinte estrutura:

WELL-99999_AAAAMMDD170106.parquet, onde:

WELL: Indica que o arquivo contém dados de um poço de petróleo real.

99999: É o número do poço, que varia de 1 a 42.

AAAAMMDD: É o ano, mês e dia em que os dados foram coletados.

O *dataset* 3W completo contém 2228 arquivos Parquet com dados de poços reais, dados simulados e dados desenhados por especialistas. Usaremos neste projeto, apenas os 1135 arquivos de poços reais sendo, 594 arquivos com observações de operações normais e 541 arquivos contendo eventos indesejados.

Pasta	**Descrição do evento**	**Identificação do Poço**	**Ano da observação**	**Qtd de arquivos**
0	Operação normal do poço	01	2017	93
0	Operação normal do poço	02	2013	8
0	Operação normal do poço	02	2017	201
0	Operação normal do poço	03	2017	26
0	Operação normal do poço	04	2014	12
0	Operação normal do poço	05	2017	81
0	Operação normal do poço	06	2017	113
0	Operação normal do poço	07	2017	2
0	Operação normal do poço	08	2017	57
0	Operação normal do poço	19	2017	1
1	Aumento abrupto de água produzida	01	2014	1
1	Aumento abrupto de água produzida	02	2014	1
1	Aumento abrupto de água produzida	06	2017	1
1	Aumento abrupto de água produzida	06	2018	1
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	02	2013	1

2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	03	2014	1
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	03	2017	1
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	03	2018	1
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	09	2017	1
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	10	2017	1
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	11	2014	13
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	12	2017	12
2	Fechamento espúrio da válvula de segurança de fundo de poço (DHSV)	13	2017	1
3	Oscilações na produção de óleo e gás	01	2017	1
3	Oscilações na produção de óleo e gás	14	2017	31
4	Instabilidade no fluxo de produção	01	2017	36
4	Instabilidade no fluxo de produção	02	2013	23
4	Instabilidade no fluxo de produção	02	2014	89
4	Instabilidade no fluxo de produção	04	2014	43
4	Instabilidade no fluxo de produção	05	2017	38
4	Instabilidade no fluxo de produção	07	2017	10
4	Instabilidade no fluxo de produção	10	2018	83
4	Instabilidade no fluxo de produção	14	2017	21
5	Perda rápida de produtividade do poço	15	2017	1
5	Perda rápida de produtividade do poço	16	2018	4
5	Perda rápida de produtividade do poço	20	2014	6
6	Restrição rápida no *choke* de produção (PCK)	02	2014	3
6	Restrição rápida no *choke* de produção (PCK)	04	2017	3
7	Formação de incrustações no *choke* de produção	01	2017	1

7	Formação de incrustações no *choke* de produção	06	2018	2
7	Formação de incrustações no *choke* de produção	21	2018	1
7	Formação de incrustações no *choke* de produção	21	2019	1
7	Formação de incrustações no *choke* de produção	22	2018	8
7	Formação de incrustações no *choke* de produção	23	2018	4
7	Formação de incrustações no *choke* de produção	24	2016	19
8	Formação de hidratos na linha de produção	19	2012	1
8	Formação de hidratos na linha de produção	19	2014	1
8	Formação de hidratos na linha de produção	19	2015	1
8	Formação de hidratos na linha de produção	19	2021	1
8	Formação de hidratos na linha de produção	25	2020	1
8	Formação de hidratos na linha de produção	26	2016	1
8	Formação de hidratos na linha de produção	26	2017	2
8	Formação de hidratos na linha de produção	27	2023	1
8	Formação de hidratos na linha de produção	28	2021	1
8	Formação de hidratos na linha de produção	29	2020	1
8	Formação de hidratos na linha de produção	30	2014	1
8	Formação de hidratos na linha de produção	31	2014	1
8	Formação de hidratos na linha de produção	32	2011	1
9	Formação de hidratos na linha de serviço	10	2018	5
9	Formação de hidratos na linha de serviço	14	2016	2
9	Formação de hidratos na linha de serviço	14	2017	1
9	Formação de hidratos na linha de serviço	15	2019	1
9	Formação de hidratos na linha de serviço	16	2015	1
9	Formação de hidratos na linha de serviço	16	2019	2
9	Formação de hidratos na linha de serviço	20	2013	1

9	Formação de hidratos na linha de serviço	33	2019	6
9	Formação de hidratos na linha de serviço	34	2019	6
9	Formação de hidratos na linha de serviço	35	2019	6
9	Formação de hidratos na linha de serviço	36	2019	3
9	Formação de hidratos na linha de serviço	37	2018	2
9	Formação de hidratos na linha de serviço	37	2019	2
9	Formação de hidratos na linha de serviço	38	2019	2
9	Formação de hidratos na linha de serviço	39	2019	2
9	Formação de hidratos na linha de serviço	40	2018	1
9	Formação de hidratos na linha de serviço	41	2018	1
9	Formação de hidratos na linha de serviço	41	2019	8
9	Formação de hidratos na linha de serviço	42	2014	5

Tabela 4 – Quantidade de arquivos disponíveis por poço, ano da observação e tipo de evento.

3.3 Configuração do Amazon S3 e Organização dos Dados

Na primeira etapa no desenvolvimento da solução de *Big Data* para detecção de anomalias em poços de petróleo, no ambiente AWS Cloud, acessamos o console do Amazon S3 e criamos um *bucket* chamado "dataset3w" na região us-east-1, para armazenar todos os dados relacionados ao projeto.

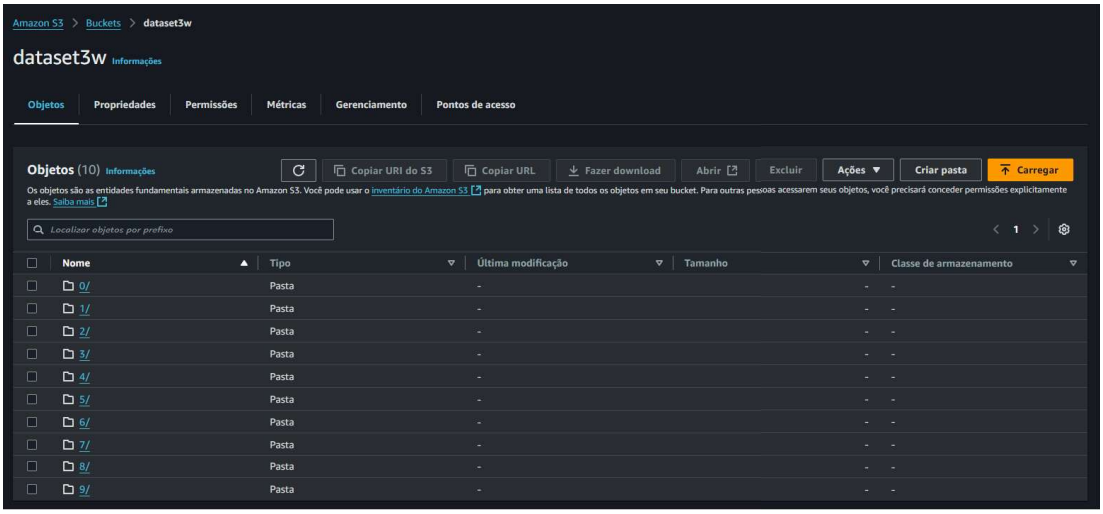

Figura 5 - Configuração inicial do *bucket* Amazon S3 no projeto. O autor, 2024.

Utilizamos o AWS CLI e o Boto3 para organizar e fazer o upload dos dados do *dataset* 3W para o *bucket* “dataset3w”.

3.4 Camada de Ingestão de Dados

Acessamos o console do Amazon Managed Workflows for Apache Airflow (MWAA) e criamos um ambiente Airflow.

- No console do MWAA, clicamos em "*Create environment*".
- Nomeamos o ambiente como airflow-dataset3w.
- Escolhemos a versão do Airflow (usada a v2.10.1 de setembro de 2024).
- Configuramos as definições de rede (VPC, subnets, security groups).
- Criamos uma role do IAM com as permissões necessárias para o Airflow acessar os serviços AWS (S3, Glue etc.).
- Configuramos o *bucket* do S3 para armazenar os DAGs e *logs* do Airflow.

1. Desenvolvimento do DAG no Airflow
   - Criamos um DAG chamado etl_dataset_3w para orquestrar as tarefas do processo ETL.

Código do DAG (etl_dataset_3w.py):

```python
from airflow import DAG
from airflow.providers.amazon.aws.operators.glue_crawler import GlueCrawlerOperator
from airflow.providers.amazon.aws.operators.glue import GlueJobOperator
from airflow.utils.dates import days_ago

default_args = {
    'owner': 'data-engineer',
    'start_date': days_ago(1),
    'email': ['kelly.castro@usp.br'],
    'email_on_failure': True,
}

with DAG(
    'etl_dataset_3w',
    default_args=default_args,
    schedule_interval='@daily',
    catchup=False,
) as dag:

    # Inicia o AWS Glue Crawler
    start_crawler = GlueCrawlerOperator(
        task_id='start_glue_crawler',
        config={'Name': 'dataset-3w-crawler'}
    )

    # Executa o job de processamento
    process_data = GlueJobOperator(
        task_id='process_data',
        job_name='process_dataset_3w_job',
        script_location='s3://scripts-bucket/process_dataset_3w.py',
        iam_role_name='GlueJobRole',
    )

    start_crawler >> process_data
```

Figura 6 - Código do DAG (etl_dataset_3w.py) no Airflow. O autor, 2024.

2. Configuração do AWS Glue Crawler

No console do AWS Glue, criamos um Crawler chamado dataset-3w-crawler.

- No console do Glue, clicamos em "Crawlers" > "Add crawler".
- Nomeamos o *crawler* como dataset-3w-crawler.
- Definimos o caminho de origem como s3://dataset-3w/.
- Criamos ou selecionamos uma *role* do IAM com permissões para acessar o S3 e operar o Glue (AWSGlueServiceRole).
- Configuramos o *crawler* para executar "*On demand*".
- Definimos o banco de dados de destino no Data Catalog como dataset_3w_db.

3. Execução do *Crawler*

Iniciamos o *crawler* manualmente ou através do Airflow (como definido no DAG). O *crawler* escaneia os dados no S3, infere os esquemas dos arquivos Parquet e atualiza o AWS Glue Data Catalog com as tabelas correspondentes.

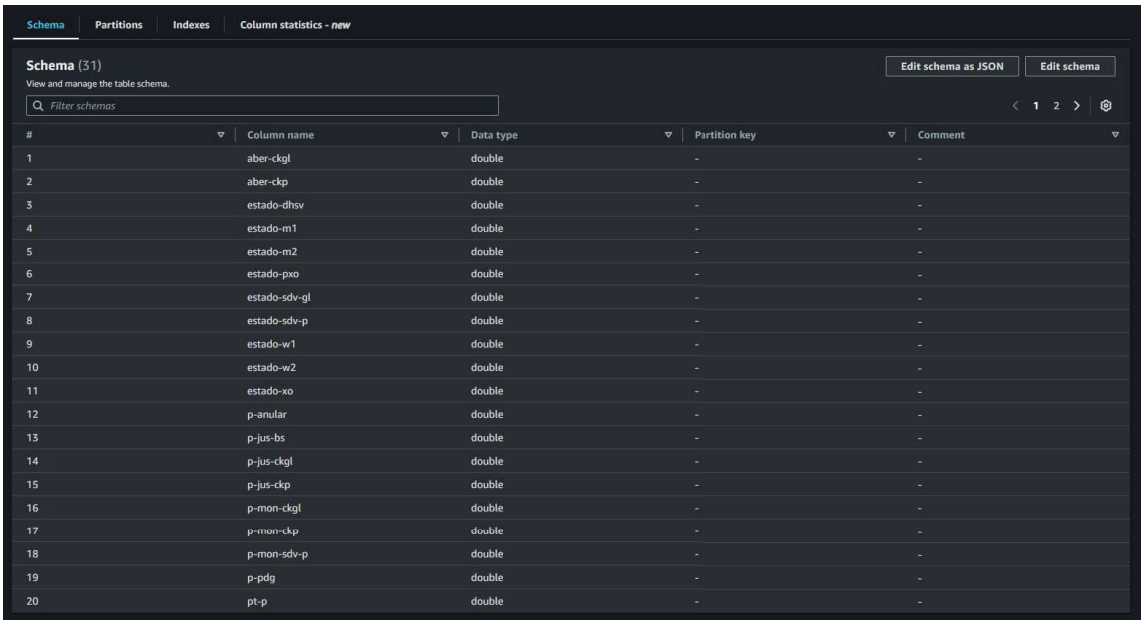

Figura 7 - Configuração do AWS Glue Crawler. O autor, 2024.

3.5 Processamento e Transformação com PySpark e BibMon

No AWS Glue, criamos um *job* chamado process_dataset_3w_job para processar os dados utilizando PySpark.

- No console do Glue, clicamos em "*Jobs*" > "*Add job*".
- Nomeamos o *job* como process_dataset_3w_job.
- Selecionamos a role do IAM com permissões necessárias.
- Escolhemos o tipo de *engine* como "Spark".

No script PySpark (process_dataset_3w.py), implementamos as etapas de transformação e modelagem dos dados.

```python
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
import bibmon  # Biblioteca BibMon
from pyspark.sql.functions import *

# Parâmetros do job
args = getResolvedOptions(sys.argv, ['process_dataset_3w_job'])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session

# Leitura dos dados catalogados
datasource0 = glueContext.create_dynamic_frame.from_catalog(
    database='dataset_3w_db',
    table_name='0',
    transformation_ctx='datasource0'
)
```

Figura 8 - Script PySpark utilizado no job process_dataset_3w_job. O autor, 2024.

No DAG do Airflow, o GlueJobOperator executa o *job* process_dataset_3w_job:

```python
process_data = GlueJobOperator(
    task_id='process_data',
    job_name='process_dataset_3w_job',
    iam_role_name='GlueJobRole',
)
```

Figura 9 - Uso do GlueJobOperator para orquestração do *job*. O autor, 2024.

3.6 Modelagem e Predição com *Machine Learning*

Acessamos o *console* do Amazon SageMaker e criamos um *notebook instance* chamado sagemaker-dataset-3w.

- No console do SageMaker, clicamos em "*Notebook instances*" > "*Create notebook instance*".
- Nomeamos a instância como sagemaker-dataset-3w.
- Escolhemos o tipo de instância (ml.t2.medium).

Selecionamos uma role do IAM com permissões para acessar o S3 (AmazonSageMakerFullAccess).

No *Jupyter Notebook*, desenvolvemos o código para treinamento dos modelos:

```python
import boto3
import sagemaker
from sagemaker import get_execution_role
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
import joblib

# Configurações iniciais
role = get_execution_role()
bucket = 'dataset-3w-processed'
prefix = '0'  # Dados de operação normal

# Listar arquivos processados no S3
s3 = boto3.resource('s3')
bucket = s3.Bucket(bucket)
files = [obj.key for obj in bucket.objects.filter(Prefix=prefix) if obj.key.endswith('.parquet')]

# Carregar os dados
data_frames = []
for file in files:
    df = pd.read_parquet(f's3://{bucket.name}/{file}')
    data_frames.append(df)

df = pd.concat(data_frames, ignore_index=True)

# Preparação dos dados
X = df.drop(['target'], axis=1)
y = df['target']

# Divisão em treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Treinamento do modelo
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Avaliação do modelo
accuracy = model.score(X_test, y_test)
print(f"Acurácia: {accuracy:.2f}")

# Salvamento do modelo
model_output_path = 's3://dataset-3w-models/random_forest_model.joblib'
joblib.dump(model, model_output_path)
```

Figura 10 - Treinamento de modelos no Amazon SageMaker. O autor, 2024.

Utilizamos o SageMaker para criar um *endpoint* de inferência:

```
from sagemaker.sklearn.model import SKLearnModel

# Definir o modelo
sklearn_model = SKLearnModel(
    model_data=model_output_path,
    role=role,
    entry_point='inference.py',  # Script de inferência
)

# Implantar o modelo
predictor = sklearn_model.deploy(
    initial_instance_count=1,
    instance_type='ml.m5.large',
    endpoint_name='rf-dataset-3w-endpoint'
)
```

Figura 11 - *Endpoint* de inferência configurado no Amazon SageMaker. O autor, 2024.

No script inference.py, definimos as funções de transformação de entrada, predição e transformação de saída.

3.7 Visualização e Análise dos Resultados

No AWS SNS, criamos um tópico chamado anomaly-alerts.

- No console do SNS, clicamos em "*Topics*" > "*Create topic*".
- Nomeamos o tópico como anomaly-alerts.
- Configuramos as permissões e políticas conforme necessário.
- Adicionamos assinaturas (*e-mail*) para receber notificações.

No Airflow, adicionamos uma tarefa que envia notificações em caso de anomalias detectadas:

```python
from airflow.providers.amazon.aws.operators.sns import SnsPublishOperator

def check_anomalies(**context):
    # Função que verifica se há anomalias nas predições
    # Retorna True se houver anomalias
    pass

notify_anomaly = SnsPublishOperator(
    task_id='notify_anomaly',
    target_arn='arn:aws:sns:region:account-id:anomaly-alerts',
    message='Anomalia detectada no poço X',
    trigger_rule='one_success',
)

process_data >> notify_anomaly
```

Figura 12 - Configuração de alertas no AWS SNS. O autor, 2024.

3.8 Monitoramento e Governança

O Airflow registra *logs* das execuções das tarefas, que podem ser acessados através da interface *web* do Airflow. Além disso, configuramos o AWS CloudWatch para monitorar os recursos e serviços.

- No console do CloudWatch, criamos alarmes para monitorar métricas como falhas nos *jobs* do Glue, utilização de *CPU* das instâncias do SageMaker, dentre outras.
- Configuramos os alarmes para enviar notificações via SNS.

Também criamos políticas e *roles* no AWS IAM para controlar o acesso aos recursos, aplicando o princípio de menor privilégio, garantindo que cada serviço tenha apenas as permissões necessárias:

- Airflow (MWAA): Acesso ao S3, Glue, SageMaker.
- AWS Glue Jobs: Acesso ao S3, Data Catalog.
- SageMaker: Acesso ao S3 para ler dados e salvar modelos.

4 CONCLUSÃO

Configuramos uma instância completa na AWS Cloud para processar o *dataset* 3W, integrando a biblioteca BibMon e utilizando diversos serviços AWS como S3, Glue, SageMaker, Airflow (MWAA) e QuickSight. Este sistema permite:

- Ingestão e Organização dos Dados: Estruturamos os dados do *dataset* 3W no S3, organizando por classes de falhas.
- Processamento e Transformação: Utilizamos o AWS Glue e PySpark com a BibMon para limpar, transformar e enriquecer os dados.
- Modelagem e Predição: Treinamos modelos de *Machine Learning* no SageMaker e implantamos como *endpoints* para inferência.
- Visualização e Análise: Acesso a consultas SQL via Athena e criação de *dashboards* para monitorar e analisar os resultados.
- Alertas e Notificações: Configuramos notificações automáticas via SNS para alertar sobre anomalias detectadas.
- Monitoramento e Governança: Implementamos práticas de segurança, monitoramento e auditoria com IAM e CloudWatch.

4.1 Contribuições do trabalho

Este projeto apresenta uma contribuição significativa ao integrar o *dataset* 3W e a biblioteca BibMon em uma instância completa de *Big Data* para o monitoramento de anomalias e a detecção preditiva de falhas em poços de petróleo. Anteriormente, esses recursos estavam disponíveis separadamente, o que dificultava a colaboração e a contribuição da comunidade *Open Source* no teste de novos modelos e na exploração das ferramentas oferecidas. Ao unificar o *dataset* e a biblioteca em uma infraestrutura integrada, baseada na Arquitetura de Referência NIST para *Big Data* (NBDRA), o projeto facilita a disseminação e o uso dos recursos, promovendo um ambiente mais propício para a inovação e o desenvolvimento de soluções avançadas.

A instância utiliza tecnologias de *Big Data* e *Machine Learning*, integrando serviços da AWS como Amazon S3, AWS Glue, Amazon SageMaker e Apache Airflow (via Amazon MWAA). Isso permite o processamento eficiente de grandes volumes de dados, a aplicação de modelos de aprendizado de máquina para a detecção de anomalias em séries temporais multivariadas e a disponibilização dos resultados por meio de visualizações interativas e alertas em tempo real. Dessa forma, o projeto não apenas aprimora a capacidade de identificar falhas em poços de petróleo, mas também estabelece uma base sólida para trabalhos futuros na área, incentivando a colaboração entre pesquisadores, engenheiros e a comunidade em geral.

4.2 Trabalhos futuros

A instância de *Big Data* proposta foi projetada para um ambiente da AWS Academy, com restrições de uso de ferramentas e funcionará de forma ainda melhor em outros ambientes de computação em nuvem com maior disponibilidade de recursos. O objetivo final é implementar a instância e comparar novos modelos de aprendizado de máquina e aprendizado profundo ao sistema, visando obter novos *benchmarks* e analisar as novas métricas de resultados.

Os resultados obtidos estão sendo gradualmente submetidos como *Pull Requests* no repositório oficial da Petrobras na plataforma GitHub e foi apresentado no *3° Workshop 3W,* promovido pela Petrobras no dia 11 de dezembro de 2024, fortalecendo a colaboração e o compartilhamento de conhecimento com a comunidade. Além disso, também sob a orientação do professor Jonas Santiago de Oliveira, da Escola Politécnica da USP, está em andamento a elaboração de um artigo científico com os resultados do projeto. A intenção é submetê-lo à *Offshore Technology Conference (OTC)*, que ocorrerá em outubro de 2025 no Rio de Janeiro. Esses esforços futuros visam ampliar o impacto do projeto, contribuindo para avanços tecnológicos na indústria de petróleo e gás e promovendo a integração entre academia e indústria.

REFERÊNCIAS BIBLIOGRÁFICAS

MELO, A., Lemos, T. S., Soares, R. M., Spina, D., Clavijo, N., Campos, L. F. D. O., ... & Pinto, J. C. BibMon: An open source Python package for process monitoring, soft sensing, and fault diagnosis. **Digital Chemical Engineering**, v. 13, p. 100182, 2024.

VARGAS, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., do Amaral, B. G., Barrionuevo, D. C., ... & Magalhães, L. P. A realistic and public dataset with rare undesirable real events in oil wells. **Journal of Petroleum Science and Engineering**, v. 181, p. 106223, 2019.

CHANG, Wo L.; BOYD, David; LEVIN, Orit. NIST big data interoperability framework: volume 6, reference architecture. 2019.

SHOIKOVA, E., Nikolov, R., Kovatcheva, E., Jekov, B., & Gotsev, L. Big Data Framework overview. **Electrotechnica & Electronica (E+ E)**, v. 55, 2020.

D'ALMEIDA, A. L., Bergiante, N. C. R., de Souza Ferreira, G., Leta, F. R., de Campos Lima, C. B., & Lima, G. B. A. Digital transformation: a review on artificial intelligence techniques in drilling and production applications. **The International Journal of Advanced Manufacturing Technology**, v. 119, n. 9, p. 5553-5582, 2022.

ANZAI, T. K., Furtado, P. H. T., de Brito, G. M., Santos, J. S., Moreira, P. C. M., Diehl, F. C., ... & Grava, W. M. Catching Failures in 10 Minutes: An Approach to No Code, Fast Track, AI-Based Real Time Process Monitoring. In: **Offshore Technology Conference Brasil**. OTC, 2023. p. D011S004R004.

MELO, A., Câmara, M. M., Clavijo, N., & Pinto, J. C. Open benchmarks for assessment of process monitoring and fault diagnosis techniques: A review and critical analysis. **Computers & Chemical Engineering**, v. 165, p. 107964, 2022.

MELO, Afrânio; CÂMARA, Maurício Melo; PINTO, José Carlos. Data-Driven Process Monitoring and Fault Diagnosis: A Comprehensive Survey. **Processes**, v. 12, n. 2, p. 251, 2024.

MORETTIN, Pedro A.; TOLOI, Clélia MC. **Análise de séries temporais, vol. 2:** Modelos multivariados e não lineares. 1. ed. Editora Blücher, 2020.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. Disponível em: https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-6-reference-architecture. Acesso em 2 de setembro de 2024.

Y. M. E. Cha Zhang. **Ensemble machine learning: methods and applications.** Springer Science & Business Media, 2012.

PETROBRAS. Projeto 3W. Disponível em: https://github.com/petrobras/3W/. Acesso em:11 de outubro de 2024.

PETROBRAS. BibMon: Biblioteca de Monitoramento de Processos. Disponível em: https://github.com/petrobras/BibMon/. Acesso em: 11 de outubro de 2024.

APACHE SOFTWARE FOUNDATION. Apache Airflow Documentation. Disponível em: https://airflow.apache.org/docs/. Acesso em: 2 de novembro de 2024.

APACHE SOFTWARE FOUNDATION. Apache Spark Documentation. Disponível em: https://spark.apache.org/docs/. Acesso em: 2 de novembro de 2024.

AMAZON WEB SERVICES. AWS Glue Documentation. Disponível em: https://docs.aws.amazon.com/glue/index.html. Acesso em: 10 de novembro de 2024.

AMAZON WEB SERVICES. Amazon S3 Documentation. Disponível em: https://docs.aws.amazon.com/s3/index.html. Acesso em: 8 de setembro de 2024.

AMAZON WEB SERVICES. Amazon SageMaker Documentation. Disponível em: https://docs.aws.amazon.com/sagemaker/index.html. Acesso em: 15 de outubro de 2024.

AMAZON WEB SERVICES. Amazon Managed Workflows for Apache Airflow (MWAA) Documentation. Disponível em: https://docs.aws.amazon.com/mwaa/index.html. Acesso em: 15 de novembro de 2024.

HDFS vs S3: Understanding the Differences, Advantages, and Use Cases - Jonas Cleveland. Disponível em: <https://jonascleveland.com/hdfs-vs-s3/>. Acesso em: 12 de outubro de 2024.

MESUT OEZDIL. AWS CloudWatch: Your Virtual Eye in the Cloud - Mesut Oezdil - Medium. Disponível em: <https://mesutoezdil.medium.com/unraveling-aws-cloudwatch-your-virtual-eye-in-the-cloud-fedb563844aa>. Acesso em: 2 de novembro de 2024.