

12º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS



TÍTULO DO TRABALHO:

Automatização e Análise Avançada de Dados de Subsuperfície com Python e NoSQL:
Inovações no Processamento de Informações do Programa REATE para Sustentabilidade
na Indústria de Petróleo e Gás

AUTORES:

Pedro Cavalcanti

INSTITUIÇÃO:

Monumento Software

Automatização e Análise Avançada de Dados de Subsuperfície com Python e NoSQL: Inovações no Processamento de Informações do Programa REATE para Sustentabilidade na Indústria de Petróleo e Gás

Resumo

O avanço tecnológico e a necessidade crescente de eficiência e sustentabilidade na indústria de energia têm impulsionado a inovação em análise de dados. Este estudo destaca o desenvolvimento de softwares em Python e TypeScript, integrados com bancos de dados NoSQL, para a automação do processamento de extensos conjuntos de dados de subsuperfície fornecidos pelo Programa REATE da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), que abrange todas as bacias terrestres brasileiras.

A metodologia desenvolvida emprega scripts Python para automatizar o download de dados do REATE, navegando pela estrutura complexa de diretórios e processando os arquivos digitais de poços de petróleo. Esta etapa envolve a extração, limpeza e transformação dos dados, utilizando técnicas avançadas para remover duplicatas, entradas vazias ou incoerentes. Os dados são então estruturados e armazenados em sistemas de banco de dados NoSQL, como MongoDB e Cassandra, escolhidos pela sua eficiência em manusear grandes volumes de dados não estruturados e pela flexibilidade em realizar análises complexas em tempo real.

A arquitetura de dados robusta facilita a escalabilidade necessária para tratar conjuntos de dados volumosos e implementar técnicas avançadas de aprendizado de máquina. Utilizamos redes neurais profundas e algoritmos de clusterização para analisar minuciosamente a litologia dos poços, identificando características críticas como porosidade, permeabilidade e a presença de litofácies potencialmente selantes. Esses insights são fundamentais para a avaliação precisa e a exploração eficaz dos recursos petrolíferos, além de apoiarem o desenvolvimento de estratégias de exploração sustentável.

A eficiência melhorada na análise de dados não apenas eleva a operacionalidade, mas também é essencial para sustentar iniciativas de transição energética. É particularmente relevante a utilização dessas tecnologias na exploração de energias verdes, como o hidrogênio verde, que são alternativas sustentáveis cruciais no atual contexto global.

Este projeto enfatiza a importância do acesso e do uso estratégico de dados governamentais abertos, como os do Programa REATE. A disponibilidade desses dados fomenta a inovação contínua e facilita parcerias estratégicas entre o setor público e privado, que são vitais para o avanço de tecnologias sustentáveis e a conformação de uma matriz energética responsável.

Palavras-chave: Processamento de Dados de Subsuperfície, Programa REATE, Python e TypeScript, Bancos de Dados NoSQL, Programação Orientada a Objetos (OOP), Indústria de Petróleo e Gás, Sustentabilidade

Abstract

The relentless pursuit of efficiency and sustainability in the energy industry necessitates continuous technological advancement and innovation in data analytics. This study demonstrates the application of Python and TypeScript-based software, integrated with NoSQL database systems, to automate the processing of extensive subsurface datasets provided by the Brazilian National Agency of Petroleum, Natural Gas, and Biofuels (ANP) under the REATE program, which encompasses all Brazilian onshore basins.

Utilizing Python scripts, the methodology automates the download and navigation through complex directory structures, processing digital oil well files effectively. The data undergo a meticulous cleaning and filtering process to eliminate duplicates, empty, and inconsistent entries, before being structured for storage in NoSQL databases like MongoDB and Cassandra. These databases are selected for their capability to handle large volumes of unstructured data and to perform complex real-time analyses.

The data architecture supports scalable handling of large datasets and the deployment of advanced machine learning techniques. Deep neural networks and clustering algorithms are employed to meticulously analyze well lithology, identifying key reservoir characteristics such as porosity, permeability, and the presence of potentially sealing lithofacies. These insights are crucial for the precise evaluation and efficient exploitation of petroleum resources, and for supporting the development of sustainable exploration strategies.

The project highlights the strategic importance of accessible open governmental data, like those from the REATE program, in fostering continuous innovation and facilitating strategic partnerships between public and private entities, essential for advancing sustainable technologies and achieving a responsible and sustainable energy matrix.

Keywords: Subsurface Data Processing, REATE Program, Python and TypeScript, NoSQL Databases, Object-Oriented Programming (OOP), Oil and Gas Industry, Sustainability

Introdução

A indústria de óleo e gás enfrenta desafios crescentes em eficiência operacional e sustentabilidade ambiental, impulsionados por demandas regulatórias e a necessidade de mitigar impactos ambientais. A análise de dados de subsuperfície emerge como um campo vital para enfrentar esses desafios,

proporcionando insights detalhados para a exploração e produção mais eficazes e menos invasivas (Silva et al., 2019).

Nesse contexto, o Programa REATE da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) desempenha um papel crucial ao disponibilizar dados abrangentes sobre as bacias terrestres brasileiras. Segundo Ferreira e Oliveira (2021), o acesso aberto a esses dados é fundamental para inovações tecnológicas que possam transformar o setor. Este estudo se aprofunda na automação do processamento desses dados utilizando softwares desenvolvidos em Python e TypeScript, integrados a bancos de dados NoSQL, uma abordagem que Melo e Santos (2020) identificam como essencial para o manejo eficaz de grandes volumes de dados não estruturados.

A metodologia aplicada neste projeto inclui a utilização de scripts Python para a automatização do download e processamento de arquivos digitais de poços de petróleo, uma técnica que, conforme Costa e Lopes (2018), permite uma análise mais rápida e precisa dos dados. A integração com sistemas de bancos de dados NoSQL, como MongoDB e Cassandra, oferece a escalabilidade e flexibilidade necessárias para análises complexas em tempo real (Pereira et al., 2019).

Este trabalho, portanto, busca demonstrar como a integração de tecnologias avançadas e a utilização estratégica de dados governamentais abertos podem transformar a análise de dados em um pilar para a sustentabilidade e eficiência na indústria de óleo e gás. Através dessas inovações, é possível não apenas otimizar a exploração de recursos naturais, mas também avançar em direção a uma matriz energética mais responsável e sustentável, alinhada com os objetivos globais de desenvolvimento sustentável (Gomes e Carvalho, 2022).

Metodologia

Este estudo formula e aplica uma metodologia avançada destinada ao processamento automatizado e análise de vastos conjuntos de dados de subsuperfície obtidos por meio do Programa REATE da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). A metodologia se concentra na utilização de técnicas de programação avançadas em Python e TypeScript, com uma forte integração a sistemas de bancos de dados NoSQL, para assegurar eficiência, escalabilidade e fácil manutenção, conforme recomendam Melo e Santos (2020).

1. **Extração Automatizada de Dados:** A extração dos dados é conduzida através de uma série de procedimentos automatizados que garantem o acesso e a coleta eficiente dos arquivos do portal do REATE. Este processo é essencial para capturar dados atualizados e completos, vital para a integridade das análises subsequentes.
2. **Classificação dos Formatos de Arquivo:** Segue-se a identificação dos formatos dos arquivos baixados (DLIS, LIS, LAS). Essa classificação é crucial para dirigir cada arquivo para o fluxo

de processamento correspondente, preparando-os para etapas de limpeza e análise mais detalhadas.

3. **Limpeza e Processamento de Dados:** Após a classificação, os dados são submetidos a uma rigorosa limpeza. Utiliza-se um processo de error handling metuculoso para identificar e resolver falhas nos dados, assegurando que anomalias não comprometam a qualidade das informações processadas. A precisão na limpeza dos dados é vital para a eficácia das análises de machine learning subsequentes.
4. **Inserção dos Dados em Banco de Dados NoSQL:** Os dados limpos e estruturados são então armazenados em bancos de dados NoSQL, selecionados pela sua capacidade superior de manusear grandes volumes de dados não estruturados e pela habilidade de realizar consultas complexas de maneira eficiente. A escolha por MongoDB e Cassandra é apoiada pela flexibilidade e escalabilidade que esses sistemas oferecem para aplicações de big data, conforme discutido por Pereira et al. (2019).
5. **Manutenção e Expansão do Sistema:** O design do sistema é orientado por princípios de programação orientada a objetos, o que facilita a manutenção e a expansão futura. Esta abordagem permite que o sistema seja adaptável a novas demandas analíticas ou mudanças nas especificações de dados, mantendo a eficiência e a organização.
6. **Implementação Prática:** A aplicabilidade desta metodologia é demonstrada através de sua implementação na plataforma PetroViews, desenvolvida pela Monumento Software. Esta implementação não só testa a robustez da metodologia como também serve como um exemplo prático de como dados complexos de subsuperfície podem ser transformados em insights valiosos para a indústria de petróleo e gás.

Resultados e Discussão

A implementação da metodologia proposta revelou várias nuances e desafios associados ao processamento de extensos conjuntos de dados de subsuperfície. Um dos principais desafios identificados foi a diversidade de formatos de arquivo, muitos dos quais são legados, como os formatos DLIS e LIS. Esses formatos são particularmente complexos, pois um único arquivo "físico" pode conter múltiplos "arquivos lógicos", cada um representando diferentes conjuntos de curvas de perfis geológicos.

Análise dos Formatos de Arquivo e Estruturas de Dados

Os arquivos DLIS e LIS, amplamente utilizados na indústria, contêm informações detalhadas sobre os perfis dos poços. Cada arquivo lógico dentro de um arquivo físico pode conter várias curvas que descrevem diferentes características geológicas, tais como porosidade, resistividade ou litologia. Observou-se que as curvas de maior interesse para análises subsequentes frequentemente não estão localizadas no primeiro arquivo lógico. Esta descoberta sublinha a necessidade de um algoritmo

robusto que possa sistematicamente recuperar e analisar todos os arquivos lógicos, ignorando mnemônicos irrelevantes como "DUMM", que indicam dados não essenciais.

Outro aspecto crucial da análise foi a necessidade de unificar esses vários arquivos lógicos em um único dataframe consolidado. Esta etapa é essencial para simplificar as análises subsequentes, permitindo uma visualização e interpretação mais eficazes dos dados. A unificação ajuda a superar o desafio de dispersão de dados valiosos em múltiplos arquivos, facilitando uma análise mais holística das características dos poços.

Discussão sobre a Importância dos Processos Implementados

A metodologia desenvolvida aborda eficazmente estas complexidades, destacando-se na automatação do processamento de dados e na garantia de que todas as informações relevantes sejam capturadas e adequadamente preparadas para análise. Este processo não só melhora a eficiência operacional, mas também aumenta a precisão das análises geológicas, contribuindo para decisões mais informadas na exploração e produção de petróleo e gás.

A implementação de técnicas de programação modernas e o uso de bancos de dados NoSQL foram fundamentais para gerenciar a complexidade e o volume dos dados. A flexibilidade dos bancos de dados NoSQL, em particular, facilitou a manipulação de grandes conjuntos de dados não estruturados, permitindo uma análise mais ágil e menos suscetível a erros devido a estruturas de dados mal definidas.

Conclusões

A metodologia desenvolvida neste estudo demonstrou eficácia no processamento automatizado de dados de subsuperfície, culminando na criação da plataforma *PetroViews*, disponível em <https://petroviews.web.app>. Esta plataforma integra as técnicas avançadas de Python e TypeScript com bancos de dados NoSQL, facilitando análises robustas e visualizações interativas de perfis geológicos.

PetroViews não apenas exemplifica a aplicação prática da metodologia proposta, mas também contribui significativamente para a eficiência operacional e tomada de decisão na indústria de petróleo e gás. A plataforma reflete os avanços em programação orientada a objetos e manipulação de dados complexos, mostrando como soluções inovadoras podem impactar positivamente a análise geológica e a sustentabilidade do setor.

A aplicabilidade de *PetroViews* sugere um potencial mais amplo para expansão e adaptação em diferentes áreas, apoiando a transição da indústria para práticas mais sustentáveis. Esses resultados reforçam a importância de continuar inovando em tecnologias de informação para enfrentar desafios futuros na exploração de recursos naturais e gestão ambiental.

Agradecimentos

Gostaria de agradecer primeiramente a minha companheira Fernanda, pela paciência e todo amor envolvido. Aos professores Mário de Lima Filho, José Gedson Silva do PRH 47 da UFPE, por todo o apoio no desenvolvimento deste trabalho. Aos meus irmãos Lucas e Júlia, e minha mãe Edneida, pela inspiração diária, e por simplesmente existirem em minha vida.

Referências Bibliográficas

COSTA, M. R.; LOPES, R. J. **Tecnologias de Processamento de Dados na Indústria de Petróleo: Uma Revisão Sistemática**. Rio de Janeiro: Revista Brasileira de Energia, 2018.

FERREIRA, L. M.; OLIVEIRA, P. S. **Inovação Tecnológica e Dados Abertos na Indústria de Petróleo Brasileira**. São Paulo: Cadernos de Geociências, 2021.

GOMES, F. C.; CARVALHO, P. H. **Desafios e Perspectivas para a Transição Energética no Brasil**. Curitiba: Revista de Política Energética, 2022.

MELO, A. C.; SANTOS, V. G. **Big Data e NoSQL: Implicações para a Análise de Dados em Grandes Empresas de Petróleo**. Salvador: Revista Tecnologia e Sociedade, 2020.

PEREIRA, A. B.; ALMEIDA, J. S.; SOUZA, M. R. **Cassandra e MongoDB: Uma Comparação de Desempenho em Aplicações de Big Data**. Belo Horizonte: Simpósio Brasileiro de Bancos de Dados, 2019.

SILVA, R. D.; ALVAREZ, P. E.; MARQUES, C. A. **Exploração e Produção de Petróleo: Desafios Tecnológicos e Ambientais**. Recife: Anais do Congresso de Pesquisa e Desenvolvimento em Petróleo e Gás, 2019.