Relatório de Progresso Primeira Entrega do Projeto de Extração de Dados de Royalties

Data: 9 de junho de 2025

Para: Observatório Social dos Royalties/PEAC

Assunto: Conclusão da Primeira Entrega - Pesquisa, Análise Técnica, Desenvolvimento de Scripts e Protótipo.

---

1. Introdução

Este relatório tem como objetivo apresentar os resultados da primeira entrega do projeto de extração de dados de royalties, conforme estabelecido na proposta inicial. A etapa compreendeu a pesquisa e análise técnica dos portais de transparência, o desenvolvimento dos scripts de extração utilizando Selenium em Python, e a apresentação de um protótipo funcional com resultados de teste.

2. Pesquisa e Análise Técnica dos Portais da Transparência

Foi realizada uma pesquisa e análise técnica detalhada dos portais da transparência das prefeituras de:

- **Aracaju/SE:** O portal da prefeitura de Aracaju foi identificado como o principal foco para o desenvolvimento do protótipo devido à complexidade e riqueza de detalhes dos dados de pagamentos, incluindo informações de "Fonte de Recurso" e "Histórico" expansíveis. A estrutura do portal exige a navegação por abas (`Pagamentos`) e a manipulação de filtros de data (Ano/Mês), além da extração de dados em tabelas dinâmicas com detalhes aninhados.

- **Barra dos Coqueiros/SE**, **Pirambu/SE e Pacatuba/SE:** Os portais de Barra dos Coqueiros, Pirambu e Pacatuba compartilham a mesma estrutura de Aracaju, com interface padronizada, sendo necessário somente alterar, via código, as URLs acessadas.

Observação: A análise de Barra dos Coqueiros, Pirambu e Pacatuba foi preliminar, focando na identificação de padrões e desafios potenciais para futuras etapas. O detalhamento do desenvolvimento e teste foi concentrado em Aracaju/SE.

3. Desenvolvimento de Scripts de Extração de Dados (Selenium com Python)

Foram desenvolvidos scripts em Python utilizando as bibliotecas `Selenium` para automação de navegação e `Pandas` para manipulação e armazenamento dos dados. O desenvolvimento incluiu as seguintes funcionalidades essenciais:

- **Inicialização e Navegação:** Função para iniciar o navegador (com opção headless) e navegar para a seção de pagamentos do portal.
- **Seleção de Período:** Funcionalidade para selecionar o ano e mês desejados nos filtros do portal.
- **Extração de Dados Multinível:** Lógica para iterar sobre as linhas da tabela principal, expandir a seção de "Mais informações" de cada registro, extrair os dados detalhados (incluindo as colunas transpostas e os campos de "Histórico Empenho" e "Histórico Pagamento") e fechar os detalhes.
- **Filtragem por Royalties:** Implementação de um filtro robusto para identificar pagamentos relacionados a royalties do petróleo, utilizando termos-chave (*‘royalty’, ‘petroleo’, ‘royalties’*) e códigos específicos de "Fonte de Recurso".
- **Paginação:** Capacidade de navegar entre as páginas da tabela de resultados para garantir a extração completa de todos os registros de um dado mês/ano.
- **Saída de Dados:** Armazenamento dos dados extraídos em formato CSV para facilitar a análise posterior.
- **Registro de Atividades (Logging):** Configuração de um sistema de logging para acompanhar o progresso e diagnosticar possíveis problemas durante a execução.

4. Protótipo e Resultados de Teste (Portal da Transparência de Aracaju/SE)

Um protótipo funcional foi desenvolvido e testado com sucesso para o portal da transparência da Prefeitura de Aracaju/SE.

Resultados de Teste:

- **Período Testado:** Mês de Setembro de 2015.
- **Acesso e Navegação:** O script demonstrou capacidade de acessar a página de pagamentos, selecionar o período desejado e aplicar os filtros corretamente.
- **Extração Detalhada:** A funcionalidade de extração conseguiu expandir os detalhes de cada registro, coletar todas as colunas da tabela principal, as informações transpostas dos detalhes (como 'Ação', 'Função', 'Fonte de Recurso', etc.) e, mais recentemente, os campos 'Histórico Empenho' e 'Histórico Pagamento'.
- **Filtragem Eficaz:** O filtro de royalties identificou corretamente os pagamentos que contêm os termos e códigos definidos na "Fonte de Recurso", garantindo que apenas os dados relevantes fossem capturados.
- **Geração de CSV:** Os dados extraídos foram compilados com sucesso em um arquivo CSV, com a estrutura de colunas acordada, facilitando a visualização e análise.
- **Modularidade:** O código foi estruturado em funções modulares, permitindo a extração sequencial para múltiplos meses/anos, e abrindo caminho para uma futura implementação de extração paralela.

A fim de validar a eficiência e escalabilidade do script de extração, vários testes foram conduzidos. Estes testes foram projetados para simular cenários de uso variados e avaliar a robustez da solução. Os seguintes testes foram realizados:

Teste 1: Mês e Ano Específicos

- **Objetivo**: Verificar a capacidade do script de extrair dados precisos para um mês e ano específicos.
- **Procedimento**: Um novo driver (instância do navegador) foi inicializado, navegando até a seção de pagamentos do portal de transparência de Aracaju/SE. Dados foram extraídos para um único mês e ano definidos.
- **Resultado**: O teste foi concluído com sucesso. Embora os dados do período selecionado tenham sido extraídos com precisão e o driver inicializado corretamente durante o teste, o tempo de execução foi significativamente alto..
    - 2025-06-04 10:00:10 - INFO - Iniciando teste para 09/2015
    - 2025-06-04 21:29:43 - INFO - Teste para 09/2015 concluído com sucesso.

Teste 2: Ano Específico com Todos os Meses

- **Objetivo**: Testar a capacidade do script de iterar por todos os meses de um ano específico, utilizando uma única instância de driver.
- **Procedimento**: O script foi configurado para percorrer todos os meses de um ano específico, mantendo o driver ativo durante todo o processo.
- **Resultado**: Este teste demonstrou a eficiência do script em lidar com múltiplas iterações e a capacidade de manter a conexão do driver sem interrupções, reduzindo o tempo de inicialização repetida. No entanto, a execução do script levou 56,7 horas (2 dias, 8 horas, 43 minutos e 36 segundos.) para finalizar a extração de todas as páginas referentes à todos os meses de um ano, evidenciando a necessidade de paralelismo na execução do código.
    - 2025-06-04 22:41:58 - INFO - Iniciando teste para o ano 2015 (todos os meses)

    - 2025-06-07 07:25:34 - INFO - Teste para o ano 2015 concluído com sucesso.

Teste 3: Teste com Vários Anos (Sequencial)

- **Objetivo**: Avaliar a função `extracao_sequencial` com uma lista personalizada de anos.
- **Procedimento**: Com base no resultado do Teste 2, a abordagem sequencial para múltiplos anos foi considerada inviável devido ao tempo excessivo de execução. Portanto, os testes subsequentes que dependiam dessa abordagem não foram realizados.

Teste 4: Execução Sequencial para Todos os Anos e Meses Definidos

- **Objetivo**: Executar a extração sequencial para todos os anos e meses definidos nas constantes `ANOS` e `MESES`.
- **Procedimento**: Mesmo caso do Teste 3.

Teste 5: Execução Paralela para Todos os Anos e Meses Definidos

- **Objetivo**: Para a próxima etapa do desenvolvimento, Implementar e testar a extração paralela usando a função `extracao_paralela`.
- **Procedimento**: A função `extracao_paralela` será utilizada, ajustando o parâmetro `max_workers` para explorar o potencial de processamento simultâneo.

Implicações para a Extração Paralela Futura

Os testes realizados fornecem evidências sólidas para justificar a implementação da extração paralela em fases futuras do projeto. O Teste 5, em particular, indica um potencial significativo para melhorias de desempenho através do processamento simultâneo de múltiplas tarefas de extração. Ao ajustar o número de `max_workers`, podemos otimizar a extração para diferentes ambientes de execução e datasets, maximizando a eficiência e reduzindo o tempo de processamento total.

Anexos:

- [Testes]
- [Dados Extraídos (Ano de 2015)]
- [Script]

5. Conclusão

A primeira entrega do projeto foi concluída com êxito, resultando em um protótipo funcional para a extração de dados de royalties do portal de transparência de Aracaju/SE. As ferramentas e metodologias empregadas demonstraram-se adequadas para o desafio, e as funcionalidades desenvolvidas comprovam a capacidade de automação e extração de dados complexos conforme as especificações.

Este protótipo serve como base sólida para a extensão do trabalho aos demais portais e para futuras fases do projeto.

---

gov.br
Documento assinado digitalmente
JEAN CLAUDIO DE SOUZA
Data: 09/06/2025 11:19:09-0300
Verifique em https://validar.iti.gov.br

Jean Claudio de Souza

Engenheiro de Computação