Relatório final de atividades

Projeto: Patente. IA: estruturação de dados de patentes para sistema de busca baseado em Inteligência Artificial

Coordenador do Projeto:	**Washington Luís Ribeiro de Carvalho Segundo**		
Coordenação:	**Coordenação Geral de Informação Científica e Técnica – CGIC**		
Nome do(a) pesquisador(a):	**Alex Sebastião Constâncio**		
Vigência da Bolsa de Pesquisa:	**06/07/2026**	**Até:**	**05/10/2026**
Encerramento em:	**05/10/2026**		
Ref. Fundação:	**32990**	**Ref. Ibict:**	**0580/2026-41**

Descrição das Atividades da Bolsa de Pesquisa:

1- Estudos para realização de protótipos do sistema de busca com aplicação de ferramentas de IA, validados para funcionalidades dos sistemas de busca (interna e externa). Meta 4- Estudos para o desenvolvimento e validação da arquitetura tecnológica das ferramentas de IA nos sistemas de busca (interna e externa). 2- Elaborar relatório técnico final de bolsa, consolidando resultados, atividades desenvolvidas e contribuições metodológicas da pesquisa.

1 INTRODUÇÃO

O projeto em questão diz respeito a uma atualização tecnológica interna do Instituto Nacional da Propriedade Industrial (INPI) para prover melhores ferramentas de consulta para público tanto interno quanto externo do INPI.

Parte desse projeto é a construção de um lago de dados (*datalake*) em Arquitetura Medalhão, que tem o objetivo de unificar todas as fontes de dados de todos os ativos do INPI em um único repositório que servirá a serviços de indexação e consulta que servirão às novas experiências de usuário em construção.

A Arquitetura Medalhão prevê a estruturação do *datalake* em três níveis progressivamente mais puros:

1. Camada bronze: dados tais quais são obtidos das diversas fontes de dados transacionais do INPI;
2. Camada prata: dados deduplicados, com eventuais diferenças de representação harmonizadas e padronizadas;
3. Camada Ouro: dados consistentes com regras de negócio próprios do INPI, para cada tipo de ativo.

Na primeira reunião de apresentação do projeto, assim como nas demais, foi esclarecido que a atuação deste bolsista se daria exatamente na Camada Ouro, especificamente sobre o ativo de Marcas.

1.1 Ambiente

O ambiente de trabalho no INPI é complexo, composto por:

1. Linguagem de programação Python 3.12;
2. Operação por meio de console VDI (Virtual Desktop Interface);
3. Comunicação por meio do Element;
4. Armazenagem do código por meio de repositórios git gerenciados por uma instância on-premisses do GitLab;
5. Esteira de CI/CD (Continuous Integration/Continuous Deploy) operacionalizada pelo Jenkins;
6. Orquestração de processos provida pelo Airflow;

7. Armazenagem de dados gerenciada pelo Apache Spark, com dados mediados pelo Iceberg;
8. Filesystem adicional provido pelo MinIO;
9. Indexação realizada por Elastic Search.

De forma geral, a construção do código necessário para construir a camada Ouro do *datalake* consiste na articulação dessas diversas ferramentas, que são mantidas por diferentes atores técnicos do INPI e Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).

2 MÉTODO

O processo técnico de produzir o segmento de Marcas da camada Ouro do *datalake* pode ser descrito abstratamente em três etapa:

1. Leitura de tabelas Iceberg componentes da camada Silver, obtidas a partir de uma sessão com o Apache Spark;
2. Aplicação de regras de negócio segundo um documento de especificação, na forma de código Python escrito dentro de um DAG do Airflow;
3. Geração de arquivos NDJSON que constituem a camada Ouro, armazenados em uma pasta específica de um *filesystem* montado via MinIO ou NFS.

O processo acima estabelece as etapas para escrita do código necessário para gerar os artefatos da camada Ouro. Esse código era armazenado no *branch* teste do repositório GitLab do ativo Marcas da camada Ouro.

Para ser colocado em atividade pelo Airflow, o DAG implementado precisa ser distribuído por uma esteira de CI/CD do Jenkins, o que acontece quando o código do *branch* teste é combinado (*merge*) com o *branch* main.

3 DESAFIOS

Apesar da arquitetura algo complexa e da ausência de definições e orientações formalmente documentadas, é possível dizer que dois foram os grandes desafios encontrados na consecução das atividades:

1. Camada Prata inexistente: segundo o Modelo Medalhão, a camada Ouro é produzida a partir de dados existentes na camada Prata, mas esta não existia para o ativo de Marcas; apenas tabelas Iceberg relativas ao ativo de Patentes existiam e estavam à disposição, prevenindo qualquer possibilidade de implementação;
2. Regras de negócio para transformação de dados inexistente: as regras de negócio fornecidas para produção da camada Ouro eram na verdade regras de comportamento para experiência do usuário, que não se prestam ao processo de transformação de dados para a confecção de um *datalake*.

Diante das duas situações acima, o trabalho se resumiu a participar de sessões de elaboração e brainstorming para superação das diversas lacunas metodológicas percebidas para dar vazão aos artefatos da camada Ouro.

Essas sessões foram dirigidas por membros sêniores do IBICT e INPI, com base nas construções existentes do ativo de Patentes, que já se encontrava em fase avançada de produção.

3 RESULTADOS ALCANÇADOS

Dadas as condições de projeto encontradas, os resultados não puderam aderir às metas previstas para a bolsa, conforme a seguir:

1. Estudos para realização de protótipos do sistema de busca com aplicação de ferramentas de IA, validados para funcionalidades dos sistemas de busca (interna e externa): na reunião de boas-vindas, fomos informados da nossa atividade na construção da camada Ouro do *datalake*, não na construção do sistema de busca;
2. Estudos para o desenvolvimento e validação da arquitetura tecnológica das ferramentas de IA nos sistemas de busca (interna e externa): não aplicável, conforme item anterior;
3. Elaborar relatório técnico final de bolsa, consolidando resultados, atividades desenvolvidas e contribuições metodológicas da pesquisa: alcançado por meio deste relatório.

No entanto, apesar das divergências operacionais experimentadas, os seguintes resultados puderam ser coletivamente alcançados, com a participação deste bolsista:

1. Elaboração da estrutura de grupos e repositórios no GitHub;
2. Padronização da estrutura de pastas de organização da camada Prata;
3. Padronização da estrutura de pastas de organização da camada Ouro;
4. Modelo de segmentação da cama Ouro por tipo de ativo e por tipo de consumidor, o que mudou a definição de passar a gerar não um, mas dois conjuntos de arquivos NDJSON;
5. Elaboração de padrões de variáveis de ambiente para escrita uniformizada e resiliente de código Python;
6. Elaboração de camada de abstração de acesso às tabelas Iceberg da camada Prata.

Essas definições promovem um grande incremento na capacidade de governança técnica do projeto do *datalake*, pois estabelecem padrões formais e concretos, pois uniformizam práticas, tornam os resultados mais previsíveis, aumentam a auditabilidade dos processos, simplificam a evolução e elevam a resiliência e escalabilidade do código.

4 CONSIDERAÇÕES FINAIS

Em função das condições de projeto encontradas, não foi possível atingir as metas da bolsa, conforme seção 3.

Diante do cenário, o propósito foi atuar para contribuir com as elaborações que faltavam para conferir ao projeto maior maturidade de governança técnica, o que foi possível por meio de sessões de reuniões com os demais membros do corpo de colaboradores.

gov.br
Documento assinado digitalmente
ALEX SEBASTIAO CONSTANCIO
Data: 22/10/2026 10:47:02-0300
Verifique em https://validar.iti.gov.br

Alex Sebastião Constâncio
Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 23/10/2026 10:06:39-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo

	Coordenador do projeto