Relatório final de atividades

Projeto: ServiçosPI: inovação e eficiência na recuperação de informações de propriedade intelectual do INPI.

Coordenador do Projeto:	**Washington Luís Ribeiro de Carvalho Segundo**		
Coordenação:	**Coordenação-Geral de Informação Científica e Técnica - CGIC**		
Nome do(a) pesquisador(a):	**Gustavo Cordeiro Galvão Van Erven**		
Vigência da Bolsa de Pesquisa:	**06/07/2026**	**Até:**	**05/10/2026**
Encerramento em:	**05/10/2026**		
Ref. Fundação:	**31849**	**Ref. Ibict:**	**0677/2024-91**

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 2- Elaborar relatório técnico final de bolsa, consolidando resultados, atividades desenvolvidas e contribuições metodológicas da pesquisa.

1 INTRODUÇÃO

O projeto “ServiçosPI: inovação e eficiência na recuperação de informações de propriedade intelectual do INPI” constitui iniciativa direcionada à modernização e aprimoramento dos serviços oferecidos pelo instituto, considerando a estrutura atual defasada e insuficiente para o atendimento das necessidades do INPI.

Para atingir o objetivo de aumentar a qualidade dos serviços por meio de maior desempenho e escalabilidade do processamento de dados, os trabalhos foram segmentados por produtos distintos, sendo eles: patentes, marcas, desenhos industriais, registros de programas de computador, topografias de circuitos integrados e indicações geográficas. Além de já ser uma divisão natural dos trabalhos do INPI, esses produtos também simplificam a divisão de tarefas e responsabilidades.

Dessa forma, em concordância com as atividades de aplicação e infraestrutura, a área de dados definiu a arquitetura de lago de dados (*data lake*) como padrão (PASUPULETI, 2015), incluindo três camadas de qualidade e acesso de dados, tendo bronze como ingestão de dados, prata com tratamentos de limpeza padronização, e a última camada, chamada de ouro, armazenando dados de maior nível de “pureza” e valor para o negócio, além de também ser a camada com aplicação das regras de negócio estabelecidas para os dados, e de acesso por aplicações, como ElasticSearch[1]. A partir dessa arquitetura foram definidas tarefas com criação de processos para carregar os dados de cada camada com suas dependências.

No período coberto por este relatório, o produto de patentes já se encontra com o processo construído, servindo de base para a definição dos demais, que se encontram atualmente na fase de geração das camadas de bronze e prata.

Com a geração de ambas as camadas, serão aplicadas as regras de negócio e criação da camada ouro, parte considerada mais complexa. Também foi definido para essa camada a entrega de dois *datasets* finais onde ocorrerem aplicações de regras de sigilo, permitindo que sistemas internos tenham acesso a todos os dados e que sistemas públicos tenham o risco de vazamento mitigado, por não possuírem informações sensíveis.

[1] https://www.elastic.co/pt/elasticsearch

Este relatório consiste no detalhamento do desenvolvimento do produto Desenhos Industriais (DI) da camada ouro.

2 MÉTODO

Para atingir os objetivos deste relatório, a pesquisa foi separada em etapas para levantamento de informações, planejamento, preparação do ambiente e implementação descritos a seguir.

I. Levantamento de Informações:

Foram revisados os documentos já produzidos no projeto e identificadas as informações relevantes para a geração do repositório *gold* de Desenhos Industriais. O escopo de identificação se estendeu das regras de negócio e dicionários de dados à infraestrutura atual e planejada para as futuras versões.

II. Planejamento:

Com as informações sobre regras de negócio, fontes de dados e infraestrutura, foi proposto um planejamento de trabalho para alinhar as necessidades de manter uma infraestrutura robusta e automatizada com um desenvolvimento mais ágil. Também foram definidas as pendências e fluxos de dados para a atividade de Desenhos Industriais. Para manter um bom alinhamento e comunicação com as equipes envolvidas, também foram agendadas reuniões fixas às segundas-feiras, no horário de 17h, além de encontros online diários, sempre que possíveleis, nesse mesmo horário.

III. Preparação do Ambiente:

A preparação do ambiente disponibilizou os diversos ambientes de trabalho já definidos desde o início do projeto, além de avançar na implantação de diversas decisões realizadas nos encontros diários, com objetivo de automatização e desenvolvimento dos novos produtos. Também é esperado que os repositórios de bronze e prata estejam disponíveis nessa etapa.

IV. Implementação:

A implementação seguirá a estrutura definida para o repositório de patentes, já com uma primeira versão pronta, além de suas atualizações decididas nas reuniões diárias. Os testes e desenvolvimentos poderão utilizar conexões diretas ao Apache Spark, mas o resultado final deve estar disponível seguindo a estrutura de Airflow/Jenkins.

3 RESULTADOS ALCANÇADOS

O levantamento de informações permitiu a compreensão da fotografia atual do projeto. Com relação à estrutura de código e documentação de infraestrutura, muitas das informações necessárias já estavam disponíveis. Entretanto, com relação às regras de negócio, os documentos ainda se encontravam em etapa de finalização, o que ocorreu recentemente.

Como ponto relevante no planejamento, foram exploradas algumas arquiteturas de fluxo de dados para automatizar a criação do *data lake*. Entretanto, essas arquiteturas dificultam o desenvolvimento por alongarem expressivamente o tempo de teste e correção (*debug*). Dessa forma, a equipe decidiu por criar uma máquina separada com Linux voltada para desenvolvimento e acesso pelo VDI, além de permitir acesso direto ao *container* do Apache Spark, simplificando a execução de código e, consequentemente, a exploração dos dados, desenvolvimento, teste e correções. Ainda assim, o produto final deve ser um repositório para disponibilização no Jenkins e execução pelo Airflow.

A partir das definições de estratégias para desenvolvimento e implantação, foram criados os ambientes necessários. O Linux está acessível e configurado com Visual Code + DevContainer para desenvolvimento do produto, além de bibliotecas para acesso ao Apache Spark. Um primeiro teste já foi realizado e está aguardando a disponibilidade da versão de produção do repositório prata.

Com a entrega das regras de negócio para Desenhos Industriais, tornou-se possível também levantar informações sobre quais tabelas serão necessárias e o planejamento das transformações e regras de sigilo a serem aplicadas. Nessa questão, foi decidido que as tabelas com regras de sigilo devem manter duas cópias, uma com a aplicação (público

externo em geral) e outra sem regras de sigilo, para uso em aplicações internas que venham a necessitar de todas as informações.

A tarefa se encontra atualmente em implementação, avaliando dados de tabelas possivelmente comuns entre Patentes e Desenhos Industriais, mas será necessário desenvolver novos códigos e, provavelmente, atualizar alguns existentes no repositório atual, para se adequar aos dados de entrada, quando disponibilizados na camada prata.

4 CONSIDERAÇÕES FINAIS

O projeto consiste em elaborar a camada ouro (gold) para os dados de Desenhos Industriais. Ela consiste em um repositório de dados, parte de um *data lake*, que mantém os dados de maior qualidade pela aplicação de diversas verificações e regras de negócio e sigilo.

As principais etapas para início da implementação já foram superadas, exceto a disponibilização dos dados da camada prata (*silver*). Entretanto, para não bloquear totalmente esta fase de construção da nova camada, as tabelas do produto Patentes estão sendo estudadas quanto ao possível reaproveitamento de tabelas.

REFERÊNCIAS

PASUPULETI, Pradeep; PURRA, Beulah Salome. **Data Lake Development with Big Data**. Birmingham: Packt Publishing, 2015. 164 p. ISBN 9781785888083

gov.br
Documento assinado digitalmente
GUSTAVO CORDEIRO GALVAO VAN ERVEN
Data: 27/10/2026 22:46:44-0300
Verifique em https://validar.iti.gov.br

Gustavo Cordeiro Galvão Van Erven

Pesquisador

De acordo:

gov.br
Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 29/10/2026 07:55:48-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo
Coordenador do projeto