UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

JOÃO PEDRO VACILOTO MONTILHA

USO DE IA GENERATIVA NA GERAÇÃO AUTOMATIZADA DE RELATÓRIOS EM PROCESSOS DE PREVENÇÃO À LAVAGEM DE DINHEIRO

CORNÉLIO PROCÓPIO

2025

versão 11.0 (abr. 25)

JOÃO PEDRO VACILOTO MONTILHA

USO DE IA GENERATIVA NA GERAÇÃO AUTOMATIZADA DE RELATÓRIOS EM PROCESSOS DE PREVENÇÃO À LAVAGEM DE DINHEIRO

USE OF GENERATIVE AI IN THE AUTOMATED GENERATION OF REPORTS IN MONEY LAUNDERING PREVENTION PROCESSES

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Software do Curso de Engenharia de Software da Universidade Tecnológica Federal do Paraná.
Orientador(a): Prof. Dr. Claiton de Oliveira

CORNÉLIO PROCÓPIO

2025

JOÃO PEDRO VACILOTO MONTILHA

USO DE IA GENERATIVA NA GERAÇÃO AUTOMATIZADA DE RELATÓRIOS EM PROCESSOS DE PREVENÇÃO À LAVAGEM DE DINHEIRO

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Software do Curso de Engenharia de Software da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 04 / novembro / 2025

Henrique Yoshikazu Shishido
Doutorado
Universidade Tecnológica Federal do Paraná

Silvio Ricardo Rodrigues Sanches
Doutorado
Universidade Tecnológica Federal do Paraná

CORNÉLIO PROCÓPIO

2025

Dedico este trabalho à minha família, pelos momentos de ausência.

AGRADECIMENTOS

Reconheço que estas breves linhas não serão suficientes para contemplar todas as pessoas que, de alguma forma, fizeram parte desta etapa tão significativa da minha vida. Desde já, peço desculpas àqueles que não estão aqui mencionados nominalmente, mas que certamente estão presentes em meus pensamentos e em minha sincera gratidão.

Ao meu orientador, Prof. Dr. Claiton de Oliveira, expresso meus mais profundos agradecimentos pela dedicação, orientação e sabedoria com que conduziu minha trajetória acadêmica.

Aos colegas de sala, pela parceria e convivência que tornaram esta caminhada mais leve e enriquecedora.

À Secretaria do Curso, pela atenção e cooperação sempre prestadas.

À minha família, registro meu reconhecimento e eterna gratidão, pois acredito que, sem o apoio, incentivo e compreensão deles, seria muito mais difícil superar os desafios que se apresentaram.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a concretização de minha formação.

RESUMO

Este trabalho apresenta o desenvolvimento e a implementação de uma solução baseada em inteligência artificial generativa para a geração automatizada de relatórios em processos de prevenção à lavagem de dinheiro (PLD). A pesquisa parte da justificativa da crescente complexidade regulatória e do elevado volume de transações financeiras que dificultam a detecção manual de operações suspeitas. O objetivo central foi propor uma arquitetura em nuvem, com base em serviços da AWS e modelos de IA generativa, capaz de processar dados estruturados em formato JSON e produzir relatórios detalhados e padronizados, seguindo diretrizes do BACEN e do COAF. A metodologia envolveu a filtragem de dados cadastrais, transacionais e contextuais, a construção de prompts otimizados para interação com a IA e a utilização de uma arquitetura serverless com processamento paralelo de múltiplos dossiês por meio do Amazon SQS. Os resultados obtidos demonstraram que o sistema é funcional, escalável e eficiente, com geração de relatórios criptografados em média de 90 segundos. Além disso, a aplicação contribui para reduzir a sobrecarga de analistas humanos, aumentando a eficiência e precisão na análise de risco. Conclui-se que a solução proposta é viável como ferramenta de apoio no compliance financeiro, embora ainda dependa da supervisão humana para evitar erros de interpretação. Como trabalhos futuros, sugere-se a integração com modelos de detecção automática de anomalias e a adequação da formatação dos relatórios ao padrão exigido pelo COAF, ampliando sua aplicabilidade em ambientes regulatórios. O fluxo de processamento emprega filas assíncronas para permitir a execução paralela de múltiplos dossiês, enquanto mecanismos de criptografia garantem a confidencialidade dos documentos gerados. Foram adotadas instruções específicas no *prompt* para evitar alucinações e para registrar como “não consultado” quaisquer dados ausentes no JSON de entrada. Os resultados demonstraram que a solução é funcional, escalável e tecnicamente eficiente, alcançando tempo médio de cerca de 90 segundos para gerar relatórios criptografados e prontos para revisão. Além de reduzir a carga operacional dos analistas, o sistema aumenta a padronização e precisão das análises, reforçando sua aplicabilidade como ferramenta de apoio ao compliance financeiro.

Palavras-chave: inteligência artificial; ia generativa; prevenção a lavagem de dinheiro.

ABSTRACT

This work presents the development and implementation of a solution based on generative artificial intelligence for the automated generation of reports in money laundering prevention (MLP) processes. The research is justified by the increasing regulatory complexity and the high volume of financial transactions, which hinder the manual detection of suspicious operations. The main objective was to propose a cloud-based architecture, using AWS services and generative AI models, capable of processing structured data in JSON format and producing detailed, standardized reports following the guidelines of BACEN and COAF. The methodology involved filtering registration, transactional, and contextual data, designing optimized prompts for interaction with AI, and implementing a serverless architecture with parallel processing of multiple dossiers through Amazon SQS. The results demonstrated that the system is functional, scalable, and efficient, generating encrypted reports in an average of 90 seconds. Furthermore, the application contributes to reducing the workload of human analysts, increasing efficiency and accuracy in risk analysis. It is concluded that the proposed solution is feasible as a support tool for financial compliance, although it still requires human supervision to avoid misinterpretations. As future work, it is suggested to integrate anomaly detection models and adapt the formatting of the generated reports to the standard required by COAF, expanding its applicability in regulatory environments. The processing flow employs asynchronous queues to allow parallel execution of multiple dossiers, while encryption mechanisms guarantee the confidentiality of the documents generated. Specific instructions were adopted in the prompt to avoid hallucinations and to record as “not consulted” any missing data in the input JSON. The results demonstrated that the solution is functional, scalable and technically efficient, achieving an average time of around 90 seconds to generate encrypted reports ready for review. In addition to reducing the operational burden on analysts, the system increases the standardization and precision of analyses, reinforcing its applicability as a tool to support financial compliance.

Keywords: artificial intelligence; generative ai; money laundering prevention.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
AWS	Amazon Web Services
BACEN	Banco Central do Brasil
CMN	Conselho Monetário Nacional
COAF	Conselho de Controle de Atividades Financeiras
IA	Inteligência Artificial
JSON	JavaScript Object Notation
KMS	Key Management Service (Amazon)
KYC	Know Your Customer
PEP	Pessoa Exposta Politicamente
PIX	Sistema de Pagamentos Instantâneos
PLD	Prevenção à Lavagem de Dinheiro
SQS	Simple Queue Service (Amazon)
TED	Transferência Eletrônica Disponível

SUMÁRIO

1 INTRODUÇÃO

A lavagem de dinheiro é um fenômeno que impacta de maneira significativa a economia mundial e nacional, permitindo que recursos oriundos de atividades ilícitas sejam inseridos no sistema financeiro com aparência de legalidade. Segundo Silva, Marques e Teixeira (2011) a lavagem de dinheiro é feita em três fases: colocação, ocultação e integração, onde, em resumo, é realizada a ingressão do dinheiro "sujo" no sistema financeiro, a ocultação de sua origem e a integração definitiva do dinheiro no sistema como "limpo", respectivamente. Esse processo não apenas fragiliza a estabilidade econômica, mas também compromete a credibilidade das instituições financeiras e a integridade do sistema regulatório.

No contexto brasileiro, o Banco Central do Brasil (BACEN), juntamente com outras entidades reguladoras, estabelece normas e diretrizes que obrigam os bancos e demais instituições financeiras a monitorar, identificar e reportar operações, além de capacitar funcionários e manter registros bancários por no mínimo 10 anos, conforme previsto na Resolução CMN nº 4.970/2021 (Brasil, 2021). Tais obrigações reforçam o papel estratégico dessas instituições na Prevenção à Lavagem de Dinheiro (PLD), ao mesmo tempo em que aumentam a complexidade e a responsabilidade dos processos de conformidade regulatória e controle interno. Estudos realizados pela Comissão Europeia (2019) indicam que a implementação de regras rígidas de PLD, embora eficazes na detecção de transações ilícitas, também impõe custos significativos de adaptação tecnológica e operacional às instituições financeiras.

Um dos principais desafios enfrentados por essas instituições é o grande volume de dados transacionais, registrando mais de 200 bilhões de movimentações financeiras no ano de 2024 (Lopes, 2025), dificultando análises rápidas e consistentes de processos manuais. Esses sistemas tradicionais de PLD resultam em um alto índice de falsos positivos — estimado em até 95% dos alertas gerados pelos sistemas de monitoramento (International, 2025; Google Cloud, 2023) — o que sobrecarrega equipes de *compliance* e gera custos operacionais elevados.

Nesse cenário, tecnologias emergentes, como a inteligência artificial (IA), têm se destacado mundialmente como ferramentas de suporte capazes de aumentar a eficiência e a precisão das análises de PLD, o que resultou em um aumento de 2 a 4 vezes na identificação efetiva de suspeitos posteriormente confirmados como envolvidos em práticas de lavagem de dinheiro em bancos como HSBC e Bradesco, enquanto diminuiu em 60% os alertas gerados (Google Cloud, 2023). Em especial, soluções que integram IA a dados estruturados permitem a geração de relatórios automatizados, contendo informações sobre clientes, transações suspeitas, entidades ou partes envolvidas e indicadores de risco, como destacado por Agarwal, Kristensen e Luget (2019). Esses dados são relevantes porque fornecem uma visão consolidada e organizada que auxilia os analistas na tomada de decisão, permitem priorizar investigações e melhoram a capacidade de identificar atividades potencialmente ilícitas de forma mais eficiente.

1.1 Justificativa

A relevância do tema é evidenciada pela crescente pressão regulatória e pela necessidade de aumentar a eficácia das práticas de *compliance* no setor financeiro. Como destaca Rodrigues (2019), o *compliance* atua como um pilar fundamental para mitigar riscos legais, financeiros e reputacionais, promovendo maior transparência e segurança nas operações das instituições. A utilização da inteligência artificial nesse contexto apresenta um grande potencial para aumentar a eficiência, reduzir custos e o tempo de trabalho dos especialistas e melhorar a qualidade das investigações (Google Cloud, 2023).

O trabalho se destaca pela inovação ao integrar técnicas de IA generativa a uma arquitetura serverless em nuvem, utilizando serviços do Amazon Web Services (AWS) como AWS Lambda, S3, SQS e Bedrock. A escolha pela IA generativa justifica-se por sua capacidade de lidar com grandes volumes de dados não estruturados e oferecer respostas contextualizadas em cenários complexos (Goodfellow; Bengio; Courville, 2016), enquanto a adoção da arquitetura serverless possibilita maior escalabilidade, resiliência e redução de custos operacionais (Jonas *et al.*, 2019), aspectos fundamentais em aplicações de PLD. Nesse sentido, a plataforma AWS apresenta-se como um ecossistema consolidado e seguro, amplamente utilizado no setor financeiro por grandes empresas como MasterCard (AWS, 2023). Além disso, o foco em relatórios automatizados pode reduzir significativamente a carga de trabalho dos analistas humanos, permitindo que se concentrem em atividades de maior valor estratégico, como a análise qualitativa e a tomada de decisão (West, 2018).

Além disso, a automatização do processo de elaboração de relatórios de PLD apresenta benefícios significativos para as instituições financeiras. Ao substituir etapas manuais repetitivas por um fluxo automatizado e padronizado, reduz-se o risco de inconsistências entre análises, minimizam-se erros decorrentes de interpretações subjetivas e acelera-se substancialmente o tempo de resposta a alertas de suspeitos (Breslow *et al.*, 2017). A automação também diminui custos operacionais associados ao trabalho intensivo de analistas, permitindo que esses profissionais concentrem seus esforços em atividades de maior valor estratégico, como investigação aprofundada, tomada de decisão e definição de políticas de risco (Everest Group, 2023). Dessa forma, a solução proposta não apenas aumenta a eficiência e a precisão do processo, mas também fortalece a capacidade institucional de lidar com grandes volumes de dados em um ambiente regulatório cada vez mais exigente.

1.2 Objetivos

Esta seção apresenta o objetivo geral e os objetivos específicos deste trabalho.

1.2.1 Objetivo Geral

Desenvolver e implementar uma solução baseada em inteligência artificial generativa para gerar relatórios automatizados de análise de risco de lavagem de dinheiro a partir de dossiês em formato JSON utilizando-se de uma arquitetura serverless.

1.2.2 Objetivos Específicos

1. Coletar e processar dados de transações bancárias em formato JSON.
2. Utilizar um modelo de IA generativa para interpretar os dados e gerar relatórios.
3. Aplicar técnicas de engenharia de prompt para otimizar a interação com modelos de IA generativa.
4. Implementar uma arquitetura escalável em nuvem utilizando AWS.
5. Permitir o processamento paralelo de múltiplos dossiês via fila SQS.
6. Validar a qualidade dos relatórios gerados com base em diretrizes do BACEN e do Conselho de Controle de Atividades Financeiras (COAF).

1.3 Delimitação do Tema

Este trabalho tem como foco a análise de dossiês previamente selecionados, após o alerta de padrões suspeitos. A solução desenvolvida baseia-se em modelos de IA generativa já disponibilizados comercialmente, sem envolver o treinamento de modelos do zero. Além disso, a implementação foi realizada exclusivamente em ambiente de nuvem AWS, não sendo objeto de comparação com outras plataformas ou provedores, apenas o detalhamento do desenvolvimento e o impacto gerado por essa aplicação.

1.4 Estrutura do Trabalho

No Capítulo 2 é apresentada a fundamentação teórica, na qual serão abordados com maior detalhamento o conceito de lavagem de dinheiro e de Prevenção à Lavagem de Dinheiro (PLD), o funcionamento dos sistemas de monitoramento de transações, a aplicação de inteligência artificial em *compliance* financeiro, modelos de linguagem, o conceito de IA generativa e as tecnologias empregadas no desenvolvimento deste trabalho.

No Capítulo 3 será descrita a metodologia adotada, com o detalhamento da arquitetura da solução, da estrutura dos dados no formato JSON, do processo de desenvolvimento do prompt para a IA, dos serviços da AWS utilizados, dos custos operacionais envolvidos, da escalabilidade do sistema e das limitações técnicas identificadas.

No Capítulo 4 são apresentados os resultados, incluindo o funcionamento do sistema, a qualidade dos relatórios gerados, o desempenho técnico do software, a estimativa de custos para a empresa e algumas limitações observadas que impactam o funcionamento do sistema automatizado.

No Capítulo 5 são discutidas as considerações finais, destacando as contribuições do trabalho, as dificuldades enfrentadas e as perspectivas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a base conceitual necessária para o desenvolvimento do trabalho, abordando aspectos relacionados à lavagem de dinheiro, sistemas de monitoramento de transações, inteligência artificial (IA) aplicada ao *compliance* financeiro, modelos de linguagem, arquitetura em nuvem e as tecnologias utilizadas na solução proposta.

2.1 Lavagem de Dinheiro e PLD (Prevenção à Lavagem de Dinheiro)

A lavagem de dinheiro é um processo pelo qual recursos obtidos de forma ilícita são introduzidos no sistema financeiro de modo a aparentar origem lícita (Aro, 2013). Tradicionalmente, esse processo é descrito em três etapas: colocação, ocultação e integração. Tanto Silva, Marques e Teixeira (2011) quanto Aro (2013) explicam que a fase de colocação consiste em inserir o dinheiro obtido ilegalmente no sistema financeiro, buscando dificultar a identificação de sua procedência. Para isso, os valores podem ser depositados em pequenas quantias, utilizados em estabelecimentos comerciais que trabalham com dinheiro em espécie, transferidos por intermediários ou registrados em nomes de terceiros, de forma a reduzir o risco de detecção.

Na etapa de ocultação, busca-se camuflar a origem dos recursos por meio de uma série de transações complexas, nacionais e internacionais, que tornam difícil o rastreamento contábil. Nessa fase, o dinheiro pode ser movimentado entre diversas contas, convertido em bens móveis ou imóveis, ou mesmo em commodities como ouro e joias (Aro, 2013). O objetivo é criar um histórico de transações que disfarce a verdadeira origem dos recursos, tornando sua rastreabilidade quase impossível para as autoridades.

Por fim, a fase de integração envolve a incorporação do capital já disfarçado ao sistema econômico, dando-lhe aparência de legalidade. Isso pode ocorrer por meio de investimentos no mercado imobiliário, financeiro ou artístico, ou através de operações simuladas que criam lucro aparente e justificativa para os recursos (Aro, 2013). Esse estágio pode ocorrer durante a fase de ocultação, já que ambas estão ligadas à movimentação do dinheiro dentro do sistema financeiro. Após essa etapa, o dinheiro pode ser utilizado livremente como se tivesse sido obtido de forma legítima, completando o ciclo da lavagem. Contudo, vale ressaltar que nem sempre a lavagem de dinheiro necessita dessas 3 fases para ser efetuada, como também há outras formas de realizar tal delito, mas ainda sim há uma grande importância em conhecer essas 3 fases para poder compreender como o processo de lavagem de dinheiro é realizado.

Para combater essa infração, foi instituída no Brasil a Lei nº 9.613/1998, complementada por resoluções do Banco Central do Brasil (BACEN) e pela atuação do COAF. A Lei nº 9.613/1998 estabelece os crimes de lavagem ou ocultação de bens, direitos e valores, além de dispor sobre a prevenção da utilização do sistema financeiro para tais atividades e criar o COAF como órgão responsável por monitorar operações suspeitas (Brasil, 1998). A norma define penas de reclusão e multa para quem pratica, auxilia ou se beneficia dessas atividades, regulamenta medidas processuais especiais, como apreensão e alienação de bens, e impõe

obrigações a pessoas físicas e jurídicas, incluindo identificação de clientes, manutenção de registros e comunicação de operações financeiras ao COAF e às autoridades competentes (Brasil, 1998). Ela também prevê a responsabilidade administrativa das instituições que não cumprirem essas normas, com sanções que vão de advertência a multa, inabilitação e cassação de autorização para funcionamento (Brasil, 1998).

2.2 Sistemas de Monitoramento de Transações

As instituições financeiras utilizam sistemas especializados para detectar atividades suspeitas em transações bancárias. Grande parte dessas ferramentas é baseada em sistemas de regras (*rule-based*), que analisam dados estruturados, como registros de transações, perfis de clientes e listas de vigilância externas, gerando alertas automáticos quando determinadas condições são atendidas (Vieira, 2018). Essas regras são formuladas com base em requisitos regulatórios, dados históricos e padrões conhecidos de crimes financeiros.

Esses sistemas frequentemente incluem um mecanismo de feedback para melhorar sua eficácia ao longo do tempo. Equipes de conformidade fornecem feedback sobre a precisão dos alertas (Vieira, 2018), que podem ser usados para refinar e atualizar as regras ou para se adequar a certos perfis de usuários.

Apesar de úteis, esses mecanismos apresentam limitações, como o alto número de falsos positivos, podendo chegar a até 95% (Google Cloud, 2023) e a necessidade constante de intervenção humana para validação. Tais fatores tornam o processo de monitoramento mais oneroso e menos eficiente.

Além disso, as instituições financeiras devem adotar um programa de *compliance* estruturado, validado pela alta administração, com diretrizes claras, governança formal (incluindo diretor de *Compliance*/PLD e auditoria interna), avaliação baseada em risco e procedimentos robustos de *know your customer* (KYC) com medidas reforçadas para clientes de maior risco, como pessoas expostas politicamente (PEP) (Intelligence, 2023). O programa também inclui treinamento contínuo, canais de denúncia, monitoramento automatizado e comunicação de operações suspeitas ao COAF, com avaliações periódicas de sua efetividade (Brasil, 2022).

2.3 Inteligência Artificial em *Compliance* Financeiro

A inteligência artificial tem se consolidado como uma alternativa para apoiar atividades de compliance, auxiliando na análise de risco, na detecção de padrões ocultos e na automação de relatórios. Ela permite a automação da monitoração de segurança, a identificação rápida de ameaças e a previsão de vulnerabilidades antes que danos irreparáveis sejam causados (Inaganti *et al.*, 2021). Além disso, a IA simplifica a gestão de *compliance*, alinhando automaticamente as configurações da nuvem aos requisitos regulatórios, como resoluções do COAF e o PCI-DSS (*Payment Card Industry Data Security Standard*), o padrão adotado internacionalmente para segurança de dados de cartões de pagamento

A IA pode ser classificada em diferentes abordagens, como a preditiva e a descritiva. De acordo com Inaganti *et al.* (2021), a IA preditiva é capaz de prever acontecimentos, como vulnerabilidades no sistema, antes que eles ocorram. Essa antecipação ocorre por meio de técnicas avançadas de *machine learning* (ML), um método de aprendizagem de IA, com dados históricos de eventos, permitindo ao sistema reconhecer padrões e os utilizando para suas previsões. Já a IA descritiva utiliza-se de algoritmos para interpretar um volume de dados e trazer informações relevantes sobre eles de acordo com seu contexto de uso (Meira, 2024).

2.4 Modelos de Linguagem e IA Generativa

A IA generativa, também conhecida como IA gen, é uma categoria de inteligência artificial capaz de gerar conteúdo original, sejam eles imagens, textos, áudio ou qualquer outro tipo de mídia (Stryker; Scapicchio, 2024). Esses modelos se baseam em aprendizagem profunda (*deep learning*, DL), sendo inspiradas em redes neurais artificiais, que buscam imitar certos aspectos do funcionamento do cérebro humano. Por meio da identificação e codificação de padrões em grandes volumes de dados, a IA generativa é capaz de compreender ao usuário e responder a comandos em linguagem natural.

A inteligência artificial generativa tem como um de seus principais avanços o desenvolvimento das *Large Language Models* (LLMs), modelos de linguagem de larga escala baseados na arquitetura *Transformer* (Zhao *et al.*, 2023). Esses modelos se caracterizam pelo enorme volume de parâmetros — frequentemente na ordem de dezenas ou centenas de bilhões —, o que possibilita um desempenho superior em tarefas de processamento de linguagem natural. O funcionamento das LLMs envolve um pré-treinamento em extensos conjuntos de dados textuais, realizado de forma auto-supervisionada, que permite a identificação de padrões linguísticos complexos e a geração de respostas em linguagem natural de forma coerente e contextualizada. À medida que a escala dos modelos aumenta, emergem capacidades adicionais, como o *few-shot learning* e o *in-context learning*, que ampliam significativamente o potencial de aplicação da IA generativa em diferentes domínios, desde assistentes virtuais até ferramentas de apoio à tomada de decisão.

Entre as LLMs mais renomadas disponíveis destacam-se o GPT-4, desenvolvido pela OpenAI, o Claude, criado pela Anthropic, o Gemini, da Google DeepMind, o DeepSeek, da chinesa DeepSeek AI, e o LLaMA, da Meta. Algumas dessas LLMs, como a LLaMA e o Claude, são integradas ao Amazon Bedrock, uma plataforma totalmente gerenciada que oferece acesso a modelos de alto desempenho, permitindo que empresas e desenvolvedores implementem soluções de IA generativa de forma escalável e segura.

2.5 Computação em Nuvem e Arquitetura *Serverless*

A computação em nuvem é um ambiente virtualizado e gerenciado de recursos computacionais, que oferece como benefícios flexibilidade, escalabilidade e disponibilidade para

aplicações e serviços (Boss *et al.*, 2007). Esse modelo possibilita que aplicações e serviços sejam acessados via internete, utilizando grandes *data centers* e servidores poderosos, acessíveis a qualquer usuário com conexão e navegador. Isso possibilita a utilização de serviços de tecnologia sob demanda, como capacidade computacional, armazenamento e bancos de dados, podendo reduzir drasticamente os custos operacionais se utilizados corretamente pela empresa (AWS, 2025). Além disso, o cliente pode utilizar apenas o serviço necessário para a aplicação, podendo ser eles infraestrutura como serviço (IaaS), plataforma como serviço (PaaS) ou software como serviço (SaaS).

A AWS é uma das principais plataformas de nuvem, oferecendo serviços variados entre as três categorias existentes de forma escalável, segura e flexível, cobrando ao usuário apenas aquilo que é utilizado pelo mesmo. Isso se deve ao fato de que a AWS disponibilizar uma arquitetura *serverless*, onde os desenvolvedores podem construir e executar aplicações sem a necessidade de gerenciar servidores, fazendo com que os desenvolvedores possam focar inteiramente no código e na lógica de negócios (Filho, 2023).

2.6 Tecnologias Específicas Utilizadas

O desenvolvimento do sistema fez uso de tecnologias da AWS, cada uma com uma função específica.

- Amazon S3: serviço de armazenamento de objetos altamente escalável da AWS, utilizado para armazenar e proteger dados de forma eficiente.
- AWS Lambda: serviço de computação *serverless* da AWS que executa código em resposta a eventos, como invocações via API Gateway ou modificações em buckets S3.
- Amazon SQS: serviço da AWS que permite a comunicação assíncrona entre componentes de sistemas distribuídos por meio de filas de mensagens.
- Amazon Bedrock: acesso a modelos de IA generativa (como o Anthropic Claude) via API, especialistas no processamento de dados e geração de *insights* a partir de grandes volumes de informação.
- API Gateway: serviço gerenciado que facilita a criação e a gestão de APIs, permitindo a criação de *endpoints* para interação com serviços *backend*.
- KMS: serviço da AWS para gerenciar chaves de criptografia.
- Secrets Manager: serviço da AWS utilizado para armazenar e gerenciar informações sensíveis, como credenciais, tokens de autenticação e configurações de sistema.

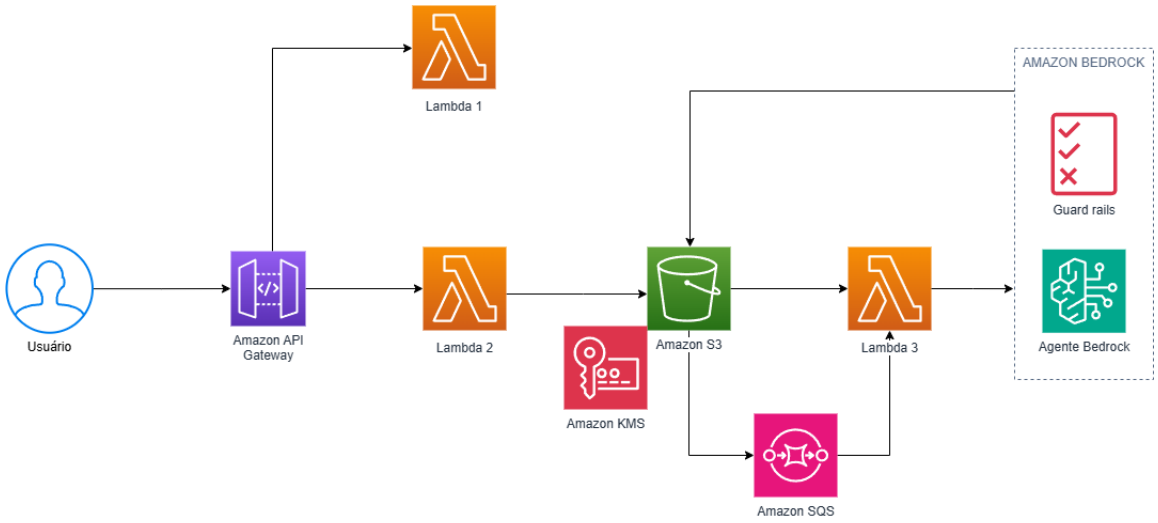
3 METODOLOGIA

Este capítulo descreve a metodologia utilizada para o desenvolvimento da solução combinando técnicas de processamento em nuvem com modelos de linguagem generativa para a geração de relatórios completos de PLD.

3.1 Arquitetura da Solução

Por questões de privacidade, não foi autorizada a apresentação do diagrama da arquitetura original da solução. Assim, a versão aqui apresentada na Figura 1 possui dados sensíveis, como URLs de *endpoints* e nomes de sistemas internos, devidamente suprimidos.

Figura 1 – Arquitetura da solução



Fonte: Elaboração própria.

3.1.1 Fluxo de Execução

3.1.1.1 Requisição Inicial e Autorização

Para iniciar o processamento, deve ser enviada uma requisição POST ao *endpoint* do API Gateway, contendo o identificador (ID) do dossiê a ser analisado como corpo, e o token de autorização (authorizationToken) como cabeçalho, como mostra o exemplo a seguir:

```
POST https://url.com/endpoint

Headers:
    authorizationToken = <token de autorização>

Body:
{
```

```
    "id": "<identificador do dossiê>"
}
```

Após a requisição, a função Lambda 1 é invocada para verificar se o evento recebido possui o campo authorizationToken como *header* e se seu valor é válido. Apesar da configuração privada dos *endpoints*, essa seria uma camada adicional de segurança, garantindo uma restrição adicional ao acesso do sistema.

3.1.1.2 Filtro dos Dados Recebidos e Retorno do *Endpoint*

Com a aprovação da requisição, o API Gateway irá invocar a função Lambda 2, cujo objetivo é realizar uma requisição GET ao *endpoint* do Microsserviço de Dossiês — um serviço externo ao projeto, mantido pela instituição responsável — e receber como resposta o conteúdo completo do dossiê em formato JSON. Em seguida, o conteúdo bruto do JSON recebido é filtrado entre dados “cadastrais”, “transacionais“ e “contextuais” essenciais ao caso. Paralelamente, os dados de bureaus de crédito são consultados no Bucket S3 e, caso existam, essas informações são combinadas aos dados previamente filtrados.

Após o processamento dos dados, o JSON resultante é salvo no Bucket S3 em uma pasta dedicada a esses dados pós filtro. Se houver uma chamada repetida para um mesmo ID, o JSON mais recente substituirá o mais antigo, que será apagado permanentemente do sistema. Após o armazenamento do arquivo no Bucket S3, a função Lambda envia uma resposta ao API Gateway uma resposta contendo o status code 200 para o cliente, ou, caso ocorra alguma falha durante este processo, ele retornará com o *status code* relacionado ao erro.

3.1.1.3 Geração do parecer

O processamento assíncrono do Parecer inicia-se com a ativação de um gatilho automático, responsável por enviar os metadados do evento, como o nome do arquivo e o bucket de origem, para a fila do Amazon SQS. A partir desse ponto, a fila aciona outro gatilho que invoca a função Lambda 3, encarregada de consumir a mensagem, acessar o arquivo no Amazon S3 e dar início à montagem do Parecer. Na sequência, o arquivo é encaminhado juntamente com um *prompt* previamente definido e a Carta Circular 4001 do Bacen, uma lista de operações e situações que indicam potenciais indícios de lavagem ou ocultação de bens, direitos e valores, bem como de financiamento ao terrorismo (Brasil, 2020) ao agente do Amazon Bedrock, que utiliza essas informações para gerar o documento completo em formato texto. Após sua criação, o Parecer é criptografado por meio de uma chave gerenciada pelo AWS KMS e armazenado novamente no Amazon S3, em uma pasta distinta.

3.2 Estrutura dos Dados (JSON)

O arquivo JSON original usado para a análise apresenta uma extensão considerável, ultrapassando 20 mil linhas. Dessa forma, tornou-se necessária a aplicação de um filtro, de modo que apenas os dados mais relevantes fossem direcionados ao bot. Os dados foram organizados em três categorias principais: (i) cadastrais, que englobam informações específicas do indivíduo investigado, como nome, idade e endereço; (ii) transacionais, referentes às operações financeiras realizadas pelo suspeito dentro do período de análise; e (iii) contextuais, que contemplam tanto informações relativas ao período em exame quanto elementos externos ao cadastro bancário, como a eventual classificação do indivíduo como PEP.

A seleção dos dados submetidos ao filtro foi realizada com o apoio de um especialista em compliance e PLD, assegurando que as informações mantidas fossem efetivamente úteis para a detecção de indícios de lavagem de dinheiro. Entre os principais dados considerados destacam-se: nome da pessoa, renda mensal declarada, volume financeiro movimentado — classificado conforme o método utilizado (PIX, cartão de crédito, depósito, transferência bancária, etc.) —, período de análise, indicação de eventual condição de PEP e histórico de bloqueios judiciais.

3.3 Desenvolvimento do *Prompt* para a IA

O *prompt* foi estruturado para orientar a geração de relatórios detalhados acerca de transações financeiras atípicas, tomando como base os dados filtrados em formato JSON e assumindo a função de um analista. O relatório deve preservar totalmente a fidelidade às informações fornecidas, sem a inserção de dados não consultados, os quais devem ser explicitamente identificados como "não consultado". Ademais, não é permitido afirmar de forma categórica a ocorrência de ilícitos, sendo aceitas apenas hipóteses cautelosas, devidamente acompanhadas de justificativas detalhadas. Também foi fornecido via *prompt* um detalhamento sobre o formato dos dados de entrada, bem como uma descrição das informações presentes em cada campo do JSON filtrado a fim de aprimorar o desempenho e a compreensão da IA em relação às informações recebidas e à tarefa a ser executada.

As instruções foram organizadas em dez partes, cada uma correspondente a uma seção específica do relatório: “Motivo da Comunicação”, “Identificação do Cliente”, “Relacionamento com o Banco”, “Reputação do Cliente”, “Período Analisado e Movimentação Financeira”, “Principais Contrapartes Identificadas”, “Sinais de Alertas Identificados”, “Consultas de Bureaus – Contrapartes”, “Conclusão e Recomendação” e “Alíneas Recomendadas”. Cada uma dessas seções contém orientações claras sobre a extração das informações no arquivo JSON e a forma adequada de apresentá-las no relatório. Além disso, foram elaborados exemplos com dados fictícios ao final do *prompt*, a fim de padronizar a geração do parecer e servir como referência para a IA.

Como parte das instruções fornecidas no *prompt*, foi estabelecida uma política explícita para lidar com informações ausentes no JSON de entrada. Sempre que um dado necessário

para a elaboração do relatório não estiver presente no arquivo original, o modelo deve registrá-lo de forma clara como “não consultado”. Essa abordagem evita inferências indevidas ou alucinações por parte da IA e sinaliza ao analista humano a necessidade de verificar se a informação realmente não existe ou se apenas não foi incluída na base de dados utilizada. Essa tratativa preserva a integridade dos resultados gerados e reforça o princípio de que a solução funciona como apoio à análise, e não como substituta da validação humana.

3.4 Configuração e Integração das Soluções AWS

- **AWS Lambda**: as três funções foram configuradas de forma idêntica, de modo que todas executassem o Python 3.13 e tivessem um tempo máximo de execução de 320 segundos.
- **API Gateway**: configurado com endpoints privados e apenas uma rota POST disponível.
- **Amazon Bedrock**: o modelo do agente selecionado foi o Claude 3.5 Sonnet e foi configurado com uma temperatura de 0,15, top P de 1, top K de 250 e o máximo de tokens na resposta de 4096.
- **Amazon SQS**: a fila foi configurada no tipo padrão, com tempo de retenção das mensagens de quatro dias, atraso de envio igual a zero segundos e tamanho máximo permitido de 256 KB por mensagem.
- **Amazon S3**: o S3 possui a configuração padrão de criação do AWS.
- **KMS**: o KMS foi configurado por uma equipe do cliente e não foram fornecidas informações sobre sua configuração.

3.5 Método de Seleção do Modelo de IA

A seleção do modelo de inteligência artificial utilizado para a geração dos relatórios foi realizada por meio de uma análise comparativa entre diferentes Large Language Models (LLMs) disponíveis no Amazon Bedrock. Esse processo envolveu a execução de uma série de testes com dossiês reais (com dados sensíveis ofuscados) com o objetivo de avaliar três critérios principais: (i) qualidade das respostas, (ii) consistência dos relatórios gerados e (iii) custo por execução.

Inicialmente, foram testados modelos como Claude 3 Opus, Claude 3.5 Sonnet, Llama 3 e Nova Pro, todos configurados com parâmetros equivalentes de temperatura e limite de tokens. Para cada modelo, foram avaliados a estrutura do relatório gerado, a aderência às instruções presentes no prompt, a capacidade de manter fidelidade aos dados fornecidos e o nível de detalhamento textual. Paralelamente, para cada execução, registrou-se o consumo de tokens e o custo aproximado associado ao uso de cada modelo.

Os testes mostraram que tanto o Claude 3 Opus quanto o Claude 3.5 Sonnet entregaram relatórios de alta qualidade, com boa contextualização e aderência às instruções do prompt. Porém, enquanto o modelo Opus apresentou desempenho ligeiramente superior, seu custo por execução era significativamente maior. Já o Claude 3.5 Sonnet manteve excelente precisão, consistência e capacidade interpretativa, com um custo muito mais baixo. Assim, o modelo selecionado foi o Claude 3.5 Sonnet, por oferecer a melhor relação entre qualidade das respostas e custo operacional.

3.6 Custos de Funcionamento do Sistema

A operação do sistema implica em custos diretamente relacionados ao consumo dos serviços da AWS. Entre os principais fatores de custo, destacam-se:

- **AWS Lambda**: cobrança por tempo de execução e quantidade de invocações.
- **API Gateway**: cobrança por requisições recebidas.
- **Amazon Bedrock**: custos mais significativos, devido ao consumo de tokens na geração de relatórios detalhados.
- **Amazon SQS**: custo proporcional ao número de mensagens enfileiradas e transferências de dados.
- **Amazon S3**: armazenamento dos arquivos JSON filtrados e pareceres.
- **KMS**: cobrança pela quantidade de operações de criptografia/descriptografia.

Embora o custo exato dependa do volume de dossiês processados, estima-se que a maior parcela esteja concentrada no uso do Amazon Bedrock, dada a alta complexidade das respostas geradas pelos modelos de linguagem.

3.7 Limitações Técnicas

- Devido à limitação de 4096 bytes por bloco de dados criptografado, o conteúdo é segmentado em múltiplos trechos, com delimitadores específicos (“|”) inseridos entre os blocos criptografados para manter a continuidade do arquivo.
- O tamanho do arquivo JSON original influencia diretamente o tempo necessário para a filtragem dos dados, de modo que quanto maior o JSON, maior será o tempo de processamento.
- Caso haja qualquer modificação nos dados ou na estrutura do JSON original, é necessário atualizar o *prompt* do agente e os dados a serem filtrados pela Lambda 1.

- Devido à utilização do modelo Claude 3.5 Sonnet nos Bedrock Agents, existe uma limitação padrão da AWS de 50 requisições por minuto para interações com esse modelo. Trata-se de uma restrição externa ao projeto; entretanto, na prática, esse limite não impacta o funcionamento da solução, uma vez que o tempo médio de geração de cada relatório (média de 90 segundos) distribui naturalmente as requisições, impedindo que o sistema atinja esse teto.
- No caso de reprocessamento de um JSON com o mesmo ID, o fluxo do sistema não será interrompido, sendo armazenado ao final apenas o resultado correspondente ao processamento mais recente.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da implementação e execução da solução proposta.

4.1 Funcionamento do Sistema

O funcionamento do sistema pode ser descrito de forma prática a partir da interação do usuário com a plataforma. Para iniciar a análise, o usuário envia uma requisição ao endpoint disponibilizado pelo API Gateway, contendo o identificador do dossiê a ser processado e um token de autorização válido. A partir desse ponto, todo o processamento ocorre de forma automática e orquestrada pelos serviços da AWS, sem necessidade de intervenção manual.

O processo completo, desde a submissão até a geração do relatório, leva em média 90 segundos, e resulta em um documento criptografado, armazenado com segurança e acessível apenas para usuários autorizados. A Figura 2 apresenta um trecho do relatório gerado pelo sistema com dados fictícios utilizado como exemplo, que pode ser consultado pelo Apêndice A.

Figura 2 – Seção 2 do relatório gerado (dados fictícios e ofuscados)

2. Identificação do Cliente
Fernando De Oliveira Batista, CPF [redacted], 33 anos, nascido em 07/03/1991. Nome da mãe: Maria Antônia De Oliveira Batista. Endereço não informado. Profissão não informada. Renda declarada de R$ 7.000,00 mensais. Renda presumida de R$ 1.550,00 (fonte Serasa, referente a 02/2022).

Participação societária:
1) IRMAOS A OBRA CONSTRUTORA E EMPREITEIRA LTDA, CNPJ [redacted], constituída em 26/06/2023, faturamento presumido anual de R$ 83.386,00. Endereço: RUA DOZE DE OUTUBRO, 15, NOSSA SENHORA APARECIDA, MANHUACU/MG, CEP 36904299.
2) VEICULOS BATISTA LTDA, CNPJ [redacted], constituída em 28/08/2024, faturamento presumido anual de R$ 82.000,00. Endereço: RUA DOZE DE OUTUBRO, 15, NOSSA SENHORA APARECIDA, MANHUACU/MG, CEP 36904299.

Vínculo empregatício:
1) WS MELO EMPREITEIRA, CNPJ [redacted], admissão em 12/03/2014, demissão em 12/09/2014.
2) IRMAOS A OBRA CONSTRUTORA E EMPREITEIRA LTDA, CNPJ [redacted], admissão em 27/06/2023, sem data de demissão informada.
3) VEICULOS BATISTA LTDA, CNPJ [redacted], admissão em 28/08/2024, sem data de demissão informada.

Informação sobre Bolsa Família não consultada.

Fonte: Elaboração própria.

> **Observação:** Os dados utilizados nesta imagem são inteiramente fictícios e foram gerados exclusivamente para fins acadêmicos e de demonstração. Qualquer semelhança com pessoas, instituições, transações ou situações reais é mera coincidência.

4.2 Relatórios Gerados

O relatório gerado pela solução segue uma estrutura padronizada em dez seções, cada uma responsável por organizar diferentes aspectos da análise. A seguir, apresenta-se um resumo de cada seção:

1. **Motivo da Comunicação** – Indica a razão que levou à geração do relatório, podendo incluir movimentações financeiras incompatíveis com o perfil declarado, operações de

alto valor ou frequência incomum, transações com contrapartes suspeitas ou qualquer outro indício de atividade atípica identificado pelas regras de monitoramento.

2. **Identificação do Cliente** – Reúne informações cadastrais, idade, vínculos empregatícios e participações societárias. São incluídos dados como nome, CPF, data de nascimento, empresas das quais o cliente participa e vínculos trabalhistas.

3. **Relacionamento com o Banco** – Descreve o histórico do cliente junto à instituição, incluindo abertura de contas, status cadastral e vínculos como representante legal.

4. **Reputação do Cliente** – Analisa possíveis indícios de risco relacionados ao cliente, como status de Pessoa Politicamente Exposta (PEP), processos judiciais, contratos públicos e menções na mídia. Quando não há dados disponíveis, registra-se explicitamente “não consultado”.

5. **Período Analisado e Movimentação Financeira** – Apresenta um resumo quantitativo das transações, incluindo totais de crédito e débito, saldo final, bem como a distribuição por tipo de operação (PIX, TED, boletos, entre outros).

6. **Principais Contrapartes Identificadas** – Lista as contrapartes mais relevantes nas operações financeiras, com valores movimentados, frequência de transações e análise de possíveis riscos, como transferências recorrentes entre familiares ou empresas de fachada.

7. **Sinais de Alerta Identificados** – Expõe indícios que justificam maior atenção, como incompatibilidade entre renda e movimentações, relações com contrapartes de risco e inconsistências geográficas nas operações.

8. **Consultas de Bureaus – Contrapartes** – Traz resultados de consultas externas sobre as contrapartes envolvidas, informando, por exemplo, se são PEPs, se possuem processos judiciais ou se foram citadas em notícias negativas.

9. **Conclusão e Recomendação** – Resume os principais achados da análise, classifica o risco (baixo, médio ou alto) e apresenta recomendações práticas, como solicitar esclarecimentos adicionais ao cliente ou comunicar o caso ao COAF.

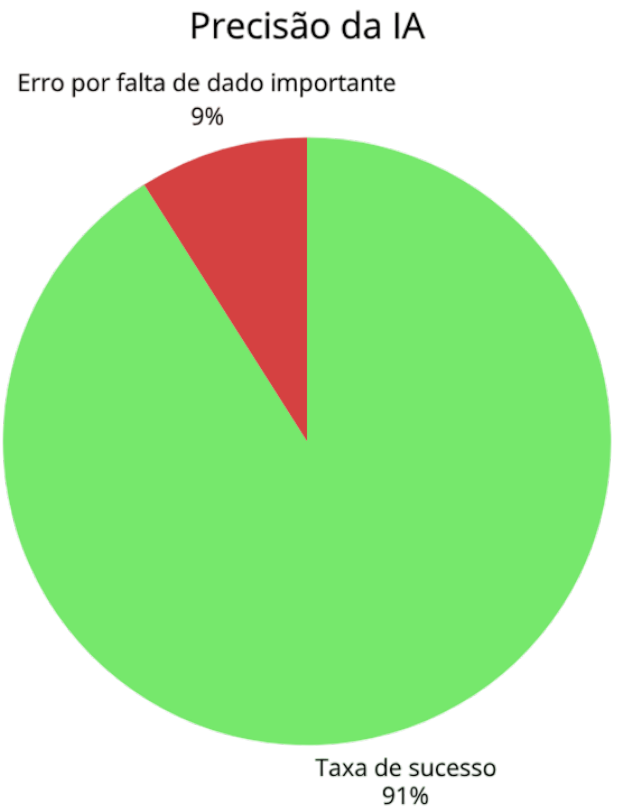
10. **Alíneas Recomendadas** – Relaciona as alíneas aplicáveis da Circular nº 4001 do Banco Central do Brasil, justificando cada recomendação de acordo com as evidências observadas no caso.

4.3 Desempenho Técnico

A solução mostrou um bom desempenho técnico, apresentando uma taxa de sucesso de 91% e um tempo médio de execução de aproximadamente 90 segundos, podendo variar de

acordo com o tamanho do dossiê processado e da necessidade de *retries*. A Figura 3 apresenta a distribuição dos resultados, destacando a proporção de relatórios gerados corretamente e os casos em que a ausência de algum dado importante impossibilitou a análise completa.

Figura 3 – Precisão da IA



Fonte: Elaboração própria.

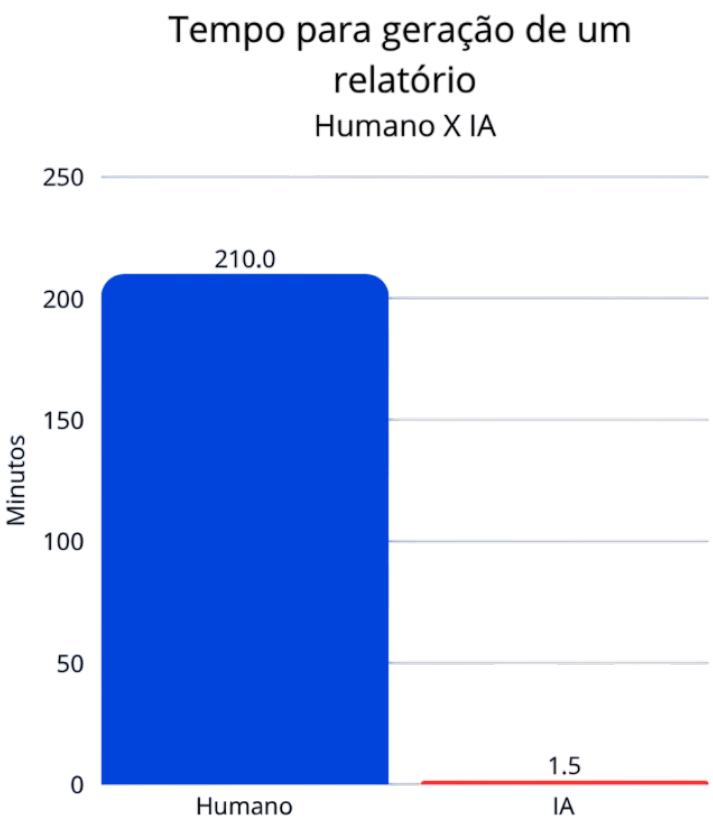
Essa performance permitiu que o tempo total de envio de um relatório ao COAF fosse reduzido para menos de uma hora, enquanto anteriormente o processo completo levava mais de 3 horas por relatório. A Figura 4 apresenta a comparação direta entre o tempo de elaboração manual e o tempo de geração automática, evidenciando o ganho expressivo de eficiência proporcionado pela automação.

Vale ressaltar que os relatórios gerados pela solução ainda precisam passar por revisão de um especialista em PLD, que também deve enquadrá-los na formatação adequada para envio, antes de serem considerados prontos.

Além disso, a arquitetura foi testada para processar até 50 dossiês simultaneamente, utilizando o Amazon SQS como mecanismo de fila, sem comprometer significativamente o tempo médio de processamento individual. Esses testes demonstraram que o sistema mantém a estabilidade e a confiabilidade mesmo sob carga elevada, distribuindo adequadamente as tarefas entre os *workers* disponíveis.

O registro de execução revelou que, em alguns casos, foi necessário realizar *retries* automáticos devido a falhas temporárias ocasionais, decorrentes da sobrecarga da IA. Apesar dessas ocorrências, a estratégia de *retry* garantiu a conclusão bem-sucedida da grande maio-

Figura 4 – Comparação de tempo: Humano x IA



Fonte: Elaboração própria.

ria dos dossiês, reforçando a robustez do sistema frente a interrupções eventuais ou picos de demanda.

4.4 Análise de Custos

O custo de utilização do sistema está diretamente relacionado à complexidade e à capacidade do modelo empregado. A maior parte do gasto é originada pelo uso do Claude 3.5 Sonnet, que permite processar grandes volumes de dados e gerar relatórios detalhados, e a grande quantidade de tokens gerados para a construção do relatório. Além disso, o custo por execução pode variar conforme a quantidade de dados analisados em cada solicitação, tornando o valor total proporcional à frequência e ao tamanho das análises realizadas. Dessa forma, os custos apresentam uma relação direta com o volume de processamento exigido, sendo possível estimar despesas futuras com base na quantidade de execuções planejadas.

Essa previsibilidade nos custos é uma vantagem do sistema em relação à análise manual. Fatores como variação na carga de trabalho, necessidade de horas extras em períodos de alta demanda ou indisponibilidade temporária de analistas podem aumentar significativamente os gastos. Dessa forma, embora mais baratos em condições normais, os custos manuais são menos constantes e mais difíceis de estimar com antecedência.

4.5 Limitações Observadas

- Os relatórios gerados por este software não substituem o trabalho do especialista, e sim servem como ferramenta para agilizar e facilitar o trabalho do mesmo.
- A qualidade dos relatórios gerados depende principalmente do JSON enviado, sendo proporcional à qualidade dos dados presentes nesse JSON e de sua estrutura.
- Ainda há o risco de ocorrer alucinações na IA, ou seja, a geração de informações que não constam nos dados fornecidos.

5 CONSIDERAÇÕES FINAIS

O presente trabalho apresentou o desenvolvimento de uma solução para apoio ao processo de Prevenção à Lavagem de Dinheiro (PLD), utilizando Inteligência Artificial e serviços em nuvem da AWS. O sistema foi capaz de demonstrar que a IA pode, de fato, ser utilizada para gerar relatórios automatizados a partir de dados estruturados, proporcionando maior eficiência no processo de análise quando comparado aos métodos exclusivamente manuais.

Os resultados obtidos mostraram que a arquitetura proposta é funcional, escalável e eficiente, permitindo o processamento paralelo de múltiplos dossiês por meio do uso do Amazon SQS e da execução de funções em AWS Lambda. A integração com o Amazon Bedrock viabilizou a geração de relatórios em linguagem natural, contendo informações úteis para analistas de *compliance*, incluindo identificação de contrapartes, sinais de alerta e alíneas do BACEN aplicáveis.

O sistema apresentou-se como uma solução prática e inovadora para o setor financeiro, principalmente pela sua capacidade de unir automação, inteligência artificial e computação em nuvem. Entretanto, também se constatou que sua eficiência depende diretamente da qualidade e estruturação dos dados de entrada, uma vez que informações incompletas ou inconsistentes podem comprometer os resultados. Além disso, vale ressaltar que a solução não possui capacidade de detectar padrões de risco por conta própria, limitando-se a interpretar e relatar os dados fornecidos.

Outro ponto importante é que, embora a IA tenha se mostrado eficaz, ainda existe o risco de erros, omissões ou interpretações equivocadas, o que reforça a necessidade de supervisão humana. Dessa forma, a solução deve ser compreendida como uma ferramenta de apoio, e não como substituta do analista especializado.

5.1 Trabalhos Futuros

Para a continuidade e evolução deste projeto, algumas melhorias podem ser exploradas:

- Integração com sistemas de detecção automática de padrões suspeitos, utilizando modelos de aprendizagem de máquina capazes de identificar anomalias em grandes volumes de transações.
- Implementação de uma interface web para facilitar o *upload* de dossiês e a visualização de relatórios por usuários não técnicos.
- Adequação da formatação dos relatórios gerados para seguir o padrão exigido pelo COAF, garantindo conformidade com as normas de reporte e facilitando a integração com sistemas regulatórios.
- Expansão da arquitetura para lidar com volumes ainda maiores de dossiês, explorando particionamento de dados e novas estratégias de paralelismo.

REFERÊNCIAS

AGARWAL, R.; KRISTENSEN, I.; LUGET, A. Como a ia generativa pode ajudar os bancos a gerir os riscos e a conformidade. **McKinsey**, „, mar. 2019.

ARO, R. Lavagem de dinheiro — origem histórica, conceito, nova legislação e fases. *In*: **Unisul de Fato e de Direito**. 3. ed. [*S.l.*]: Unisul, 2013. p. 167–177.

AWS. **AWS para serviços financeiros**, . 2023. Disponível em: https://aws.amazon.com/pt/solutions/case-studies/mastercard-ai-ml-testimonial. Acesso em: 11 set. 2025.

AWS. **O que é computação em nuvem?**, . 2025. Disponível em: https://aws.amazon.com/pt/what-is-cloud-computing/. Acesso em: 03 set. 2025.

BOSS, G. *et al.* IBM **Cloud Computing**. [*S.l.*], 2007. Disponível em: http://www.ibm.com/developerworks/websphere/zones/hipods/. Acesso em: 03 set. 2025.

BRASIL. **Carta Circular nº 4.001, de 29 de agosto de 2020**. BACEN, 2020.

BRASIL. **Lei nº 9.613, de 3 de março de 1998**, . 1998. Dispõe sobre os crimes de lavagem de dinheiro e dá outras providências.

BRASIL. **Resolução COAF nº 41, de 8 de agosto de 2022**. Brasília COAF, 2022. Acesso em: 31 ago. 2025. Disponível em: https://www.gov.br/coaf/pt-br/acesso-a-informacao/Institucional/a-atividade-de-supervisao/regulacao/supervisao/normas-1/resolucao-coaf-no-041-de-08-08.2022.

BRASIL. **Resolução nº 4.970, de 25 de novembro de 2021**. BACEN, 2021.

BRESLOW, S. *et al.* Article **New analytical tools and surgical automation can help banks take the fight to fraudsters.**, . 2017. Disponível em: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-new-frontier-in-anti-money-laundering.

EUROPEIA, C. Report on the costs of compliance with anti-money laundering regulations. **EBA Research Papers**, „, ago. 2019. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52019DC0360. Acesso em: 30 ago. 2025.

Everest Group. Everest Group **Financial Crime and Compliance: Automation Trends and Cost Reduction Benefits**. [*S.l.*], 2023. Acesso em: 15 nov. 2025. Disponível em: https://www.everestgrp.com/.

FILHO, R. F. N. **Arquitetura Serverless**. 2023. Dissertação (Trabalho de Conclusão de Curso) — Pontifícia Universidade Católica de Goiás 2023. Disponível em: https://repositorio.pucgoias.edu.br/jspui/handle/123456789/5925. Acesso em: 03 set. 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [*S.l.*]: MIT Press, 2016. http://www.deeplearningbook.org.

Google Cloud. **Google Cloud Launches AI-Powered Anti Money Laundering Product for Financial Institutions**, . 2023. Disponível em: https://www.googlecloudpresscorner.com/2023-06-21-Google-Cloud-Launches-AI-Powered-Anti-Money-Laundering-Product-for-Financial-Institutions
Acesso em: 02 set. 2025.

INAGANTI, A. C. *et al.* Cloud security posture management (cspm) with ai: Automating compliance and threat detection. **Artificial Intelligence and Machine Learning Review**, v. 2, n. 4, p. 8–18, 2021. Disponível em: https://doi.org/10.69987/.

INTELLIGENCE, A. **AML/CTF compliance in Brazil**, . 2023. Disponível em: https://arctic-intelligence.com/countries/compliance-brazil. Acesso em: 31 ago. 2025.

INTERNATIONAL, R. B. **How high false positives AML rate hurt banks, fintechs, customers**, . 2025. Disponível em: https://www.retailbankerinternational.com/comment/hidden-cost-of-aml-how-false-positives-hurt-banks-fintechs-customers/. Acesso em: 30 ago. 2025.

JONAS, E. *et al.* Cloud programming simplified: A berkeley view on serverless computing. **arXiv, Cornell University**, „, 2019. Disponível em: https://arxiv.org/abs/1902.03383.

LOPES, F. **Mobile Banking é responsável por 75bancárias no Brasil**, . 2025. Disponível em: https://febrabantech.febraban.org.br/temas/inovacao/mobile-banking-e-responsavel-por-75-das-transacoes-bancarias-no-brasil. Acesso em: 30 ago. 2025.

MEIRA, S. **Inteligência Artificial, Estratégia e Diferenciais Competitivos Sustentáveis**, . 2024. Disponível em: https://pt.scribd.com/document/770264147/Inteligencia-Artificial-Estrategia-e-Diferenciais-Competitivos-Sustentaveis. Acesso em: 02 set. 2025.

RODRIGUES, J. V. F. **Compliance nas Instituições Bancárias**: Relação com o sistema de controlo interno e a auditoria interna. 2019. Dissertação (Mestrado) — Universidade Católica Portuguesa 2019. Disponível em: https://core.ac.uk/download/pdf/322889820.pdf. Acesso em: 31 ago. 2025.

SILVA, J. L. R. da; MARQUES, L. F. B.; TEIXEIRA, R. Prevenção à lavagem de dinheiro em instituições financeiras: avaliação do grau de aderência aos controles internos. **BASE - Revista de Administração e Contabilidade da UNISINOS**, v. 8, n. 4, p. 300–310, 2011. Disponível em: https://www.redalyc.org/articulo.oa?id=337228648004. Acesso em: 30 ago. 2025.

STRYKER, C.; SCAPICCHIO, M. IBM **O que é a IA generativa?**, . 2024. Disponível em: https://www.ibm.com/br-pt/think/topics/generative-ai. Acesso em: 02 set. 2025.

VIEIRA, V. L. R. A atuação do coaf na prevenção à lavagem de dinheiro à luz da teoria da regulação responsiva. **Revista de Direito Setorial e Regulatório**, Brasília„, 2018.

WEST, D. M. **The Future of Work: Robots, AI, and Automation**. Washington, D.C.: Brookings Institution Press, 2018. ISBN 9780815732945.

ZHAO, W. X. *et al.* A survey of large language models. **arXiv preprint**, arXiv:2303.18223„, 2023. Disponível em: https://arxiv.org/pdf/2303.18223. Acesso em: 03 set. 2025.

APÊNDICE A – Relatório Gerado

Observação: Os dados utilizados neste relatório são inteiramente fictícios e foram gerados exclusivamente para fins acadêmicos e de demonstração. Qualquer semelhança com pessoas, instituições, transações ou situações reais é mera coincidência.

Relatório de Comunicação de Transações Atípicas

1. Motivo da Comunicação

A transação realizada pelo cliente Fernando De Oliveira Batista, CPF ***.***.***-**, foi sinalizada por movimentação financeira possivelmente incompatível com o perfil declarado. O evento ocorreu em 14/02/2025 e a regra acionada foi *"Movimentações aparentemente incompatíveis com o perfil"*.

O cliente possui renda mensal declarada de R$ 7.000,00, mas movimentou R$ 710.886,70 em entradas e R$ 770.336,60 em saídas em um período de 6 meses, o que parece ser incompatível com seu perfil financeiro.

2. Identificação do Cliente

- Nome: Fernando De Oliveira Batista
- CPF: ***.***.***-**
- Idade: 33 anos, nascido em 07/03/1991
- Nome da mãe: Maria Antônia De Oliveira Batista
- Endereço: não informado
- Profissão: não informada
- Renda declarada: R$ 7.000,00 mensais
- Renda presumida: R$ 1.550,00 (fonte Serasa, 02/2022)

Participação societária:

1. IRMÃOS A OBRA CONSTRUTORA E EMPREITEIRA LTDA, CNPJ **.***.***/****-**, constituída em 26/06/2023, faturamento presumido anual de R$ 83.386,00. Endereço: Rua Doze de Outubro, 15, Nossa Senhora Aparecida, Manhuaçu/MG, CEP 36904-299.
2. VEÍCULOS BATISTA LTDA, CNPJ **.***.***/****-**, constituída em 28/08/2024, faturamento presumido anual de R$ 82.000,00. Endereço: Rua Doze de Outubro, 15, Nossa Senhora Aparecida, Manhuaçu/MG, CEP 36904-299.

Vínculo empregatício:

1. WS Melo Empreiteira, CNPJ **.***.***/****-**, admissão em 12/03/2014, demissão em 12/09/2014.
2. IRMÃOS A OBRA CONSTRUTORA E EMPREITEIRA LTDA, CNPJ **.***.***/****-**, admissão em 27/06/2023, sem data de demissão.
3. VEÍCULOS BATISTA LTDA, CNPJ **.***.***/****-**, admissão em 28/08/2024, sem data de demissão.

Informação sobre Bolsa Família não consultada.

3. Relacionamento com o Banco

O cliente mantém relacionamento com o Banco desde 24/07/2022.

- Conta de pagamento (ID 80454482), titular, status “STANDARD”.
- Representante legal de conta de pagamento (ID 6701efb55d7c49f0aa0c51c2) desde 05/10/2024.

4. Reputação do Cliente

- PEP: não consultado.
- Doações eleitorais: não consultado.
- Contratos públicos: não consultado.
- Processos judiciais: não foram encontrados.
- Reportagens relevantes: não foram encontradas.

5. Período Analisado e Movimentação Financeira

Período: 18/08/2024 a 14/02/2025 (6 meses)

- Crédito: R$ 710.886,70
- Débito: R$ 770.336,60
- Total: R$ 1.481.223,30
- Saldo final: R$ 17,18 (em 14/02/2025)

Distribuição por meio:

- Crédito: R$ 535.046,48 (75,26%) via PIX (32 transações)
- Crédito: R$ 175.800,00 (24,73%) via TED (2 transações)
- Crédito: R$ 40,22 (0,01%) via outros meios (6 transações)
- Débito: R$ 674.687,74 (87,58%) via PIX (57 transações)
- Débito: R$ 90.200,00 (11,71%) via TED (1 transação)
- Débito: R$ 5.449,23 (0,71%) via outros meios (14 transações)

6. Principais Contrapartes Identificadas

Crédito:

- R$ 151.900,00 (21,37%) – 2 transações via PIX de Natanael Rodrigues Gomes (CPF ***.***.***-**).
- R$ 133.300,00 (18,75%) – 1 transação via TED de Ledir José de Souza (CPF ***.***.***-**).
- R$ 77.887,00 (10,96%) – 2 transações via PIX de Veículos Batista LTDA (CNPJ **.***.***/****-**).
- R$ 68.000,00 (9,57%) – 3 transações via PIX de Fernando Oliveira Batista (CPF ***.***.***-**).
- R$ 58.000,00 (8,16%) – 3 transações via PIX de Fernanda de Oliveira Batista (CPF ***.***.***-**).
- R$ 57.175,00 (8,04%) – 10 transações via PIX de Fernando de Oliveira Batista (CPF ***.***.***-**).
- R$ 46.591,48 (6,55%) – 2 transações via PIX de M13 Café (CNPJ **.***.***/****-**).

Débito:

- R$ 412.700,00 (53,59%) – 3 transações via PIX para Localiza Rent a Car SA (CNPJ **.***.***/****-**).
- R$ 225.000,00 (29,22%) – 2 transações via PIX para Companhia de Locação das Américas (CNPJ **.***.***/****-**).

7. Sinais de Alerta Identificados

- Movimentações possivelmente incompatíveis: volume de transações superior à renda declarada.
- Relação com contrapartes de risco: transações de alto valor com empresas de aluguel de veículos.
- Análise geográfica: operações em Manhuaçu/MG e Belo Horizonte/MG sem justificativa clara.

8. Consultas de Bureaus – Contrapartes

Natanael Rodrigues Gomes (CPF *.***.***-**):** PEP: não é PEP. Processos judiciais: não encontrados. Mídia negativa: não encontrada.

Localiza Rent a Car SA (CNPJ **.*.***/****-**):** Informações não consultadas.

9. Conclusão e Recomendação

Resumo: movimentação superior à renda declarada, transações de alto valor com locadoras de veículos sem justificativa econômica clara, indicando possível atividade ilícita.

Classificação: Alto risco.

Recomendações:

1. Solicitar esclarecimentos ao cliente sobre origem dos recursos e transações com locadoras.
2. Considerar comunicação ao COAF devido às movimentações atípicas.

10. Alíneas Recomendadas

A. IV-a) (Código 1045): Movimentação de recursos incompatível com patrimônio, atividade econômica ou ocupação profissional. *Justificativa:* movimentou R$ 1.481.223,30 em 6 meses, valor muito superior à renda declarada.
B. IV-c) (Código 1047): Movimentação de recursos de alto valor, de forma contumaz, em benefício de terceiros. *Justificativa:* transferências PIX elevadas para locadoras, totalizando R$ 637.700,00.
C. IV-e) (Código 1049): Movimentação significativa em conta até então pouco movimentada. *Justificativa:* depósitos expressivos via PIX e TED, incluindo R$ 133.300,00 de uma única fonte.
D. IV-m) (Código 1057): Contas com créditos e débitos via instrumentos não característicos para ocupação do cliente. *Justificativa:* transações de alto valor com empresas de aluguel de veículos.
E. IV-n) (Código 1058): Recebimento de depósitos de diversas origens sem fundamentação econômico-financeira. *Justificativa:* depósitos de múltiplas fontes sem justificativa compatível com perfil.