NEXOBANK - SOLUTION DESIGN DOCUMENT
Projeto: Módulo de Insights Financeiros com IA Generativa
Versão: 1.0
Data: 29/07/2025
Autor: Jhonat Heberson Avelino de Souza

Sumário

Revisores:

X

João Andrade
CTO NrxoBank

X

Lucas Freitas
Especialista em Segurança da Informação

X

Ana Paula
Gerente de Engenharia de Dados

O presente documento descreve a proposta arquitetural para o novo módulo de insights financeiros do NexoBank, com foco no uso de Inteligência Artificial Generativa de forma segura, reprodutível e integrada ao ambiente multicloud já adotado pela organização.

O NexoBank é uma fintech digital que já utiliza modelos tradicionais de machine learning, e agora deseja ampliar suas capacidades para gerar insights personalizados a partir dos dados dos clientes, combinando transações financeiras, contratos e documentos internos. Esse novo módulo será acessado exclusivamente via o aplicativo mobile dos clientes Private e Select.

Para dar suporte a esse novo recurso, propomos uma arquitetura baseada em serviços gerenciados da AWS e integração com o legado existente na Oracle Cloud (OCI). A conectividade multicloud entre a Oracle Cloud Infrastructure (OCI) e a Amazon Web Services (AWS) usando o backbone Megaport, O Backbone Megaport mostra que um Roteador de Nuvem Megaport interconecta o VXC que anexa ao Oracle Cloud Infrastructure FastConnect e aos circuitos VIF Privados do AWS Direct Connect, A região do OCI mostra uma Rede Virtual na Nuvem (VCN) com o serviço de nuvem implantado. A VCN usada para o banco de dados tem um gateway de roteamento dinâmico (DRG). O serviço de nuvem se conecta ao VXC dentro do Backbone Megaport usando FastConnect por meio do gateway de roteamento dinâmico. Região da AWS mostra a Nuvem Privada Virtual (VPC), Na região da AWS, a VPC usada para acessar API Gateway tem um gateway virtual (VGW). Um VIF privado do AWS Direct Connect é usado para associar o VGW ao AWS Direct Connect (Oracle, s.d.). O diagrama abaixo ilustra a visão macro da solução.

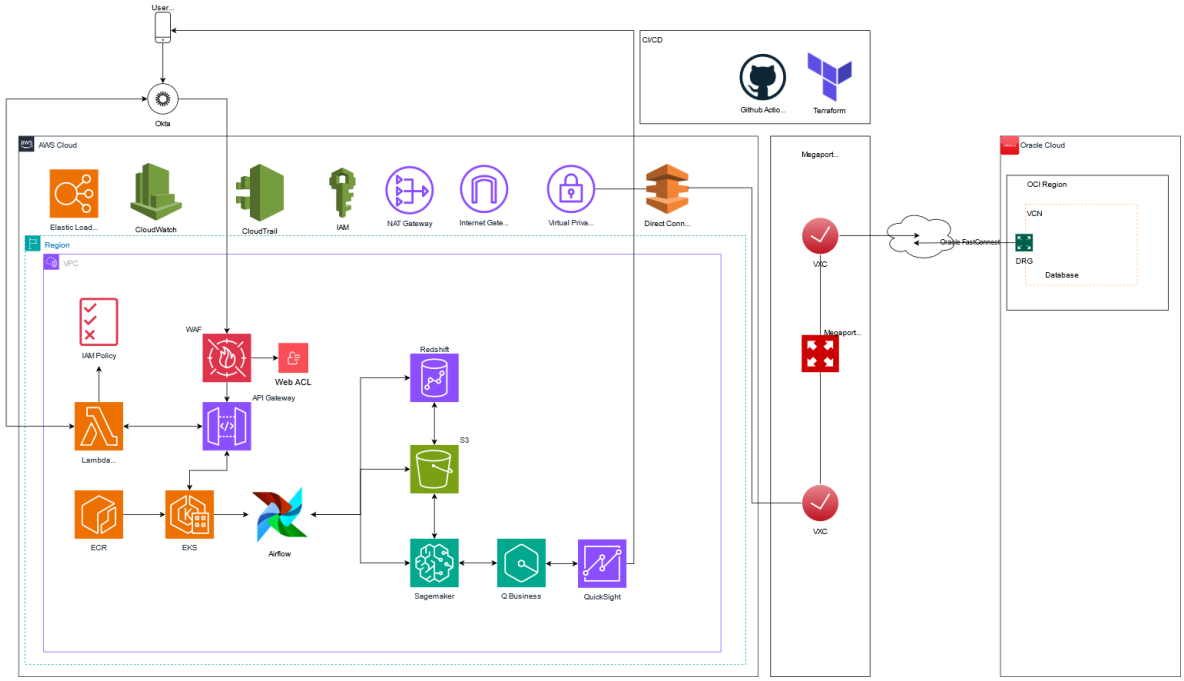

Figure 1: Diagrama de Arquitetura - Nível de Contexto

Arquitetura Proposta

A arquitetura contempla três grandes blocos: o aplicativo mobile, o backend de integração e os serviços de inteligência artificial e analytics. O aplicativo, desenvolvido em React Native, autentica os usuários via Auth0 utilizando OAuth2 e JWT. Okta é ativado, criando uma solicitação de token web JSON (JWT) para entrega de OTP por meio do Amazon API Gateway. O AWS WAF protege o endpoint do API Gateway aplicando regras gerenciadas pela AWS para bloquear tráfego malicioso. Todo o tráfego é filtrado pelas listas de controle de acesso (ACL) da Web do AWS WAF , e as solicitações consideradas seguras podem passar pelo API Gateway . O API Gateway primeiro recebe a solicitação JWT do Okta. Em seguida, ele invoca uma função personalizada do AWS Lambda que atua como um autorizador para validar o token JWT antes de permitir que a solicitação prossiga. O autorizador Lambda é responsável por verificar a integridade e a validade do token JWT. Ele realiza diversas verificações para garantir que o token seja válido. O autorizador do Lambda verifica o token JWT decodificando-o, usando a chave pública do Okta para validar a assinatura e verificando o tempo de expiração. Se o token JWT for válido, o autorizador do Lambda cria uma política do AWS Identity and Access Management (IAM) que concede permissão para invocar o API Gateway. O autorizador do Lambda retorna a política do IAM para o API Gateway. Se o acesso for permitido, o API Gateway é invocado e encaminha a solicitação para backend em Node.js (AWS, 2025), que encaminha as chamadas para os serviços apropriados.

Na nuvem AWS, o Amazon API Gateway roteia as solicitações para backend, que por sua vez acionar o orquestrador e acessam os dados no Amazon Redshift,

invocam processamentos no SageMaker e consultam a IA generativa via Amazon Q Business. Documentos regulatórios armazenados na OCI são integrados ao pipeline de embeddings através do Apache Airflow, garantindo atualização constante da base de conhecimento. Para a visualização dos insights, utiliza-se o Amazon QuickSight com dashboards personalizados, embutidos no app mobile.

Abaixo, o diagrama visual representa essa arquitetura.

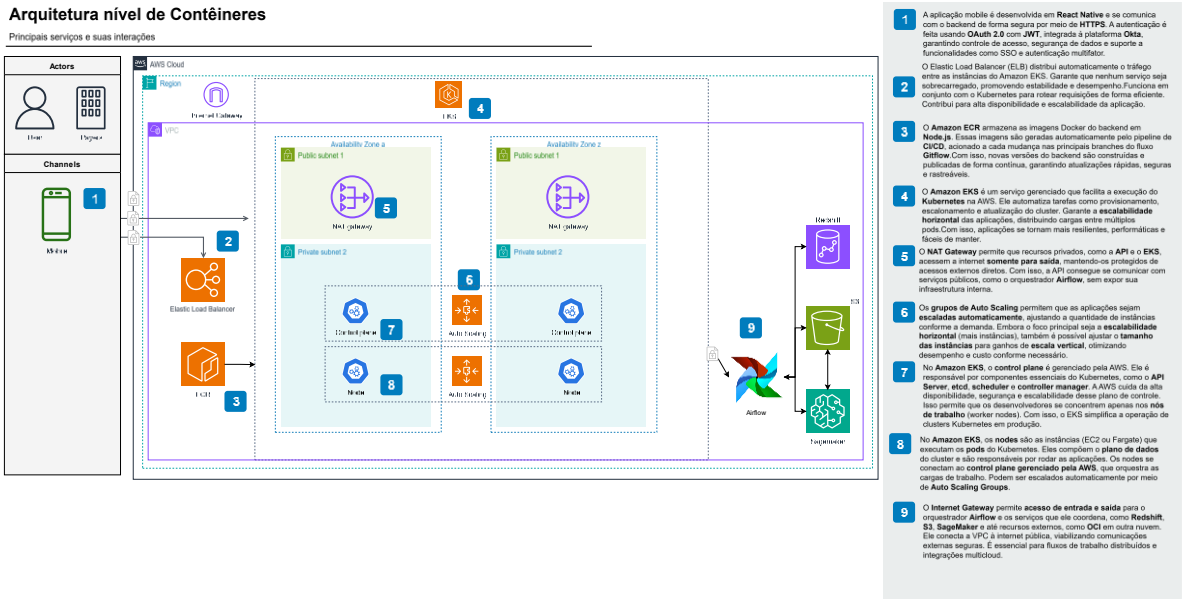

Figure 2: Diagrama de Arquitetura - Nível de Contêineres

Estratégia de IA Generativa

A abordagem central da solução é baseada em RAG (Retrieval-Augmented Generation). Por meio dessa técnica, documentos internos (como PDFs, XMLs e contratos) são vetorizados e armazenados como embeddings. Esses vetores são utilizados para enriquecer os prompts enviados ao Amazon Q Business, que gera respostas contextuais e personalizadas para cada cliente.

A execução desse processo envolve uma esteira de dados orquestrada com Apache Airflow, que coleta e transforma dados tanto da AWS quanto da Oracle Cloud. O modelo gerador, hospedado no Q Business, recebe entradas compostas por informações transacionais, perfis de cliente e conteúdos regulatórios, e devolve insights em linguagem natural.

Para garantir a reprodutibilidade, cada insight gerado é registrado com um identificador único, contendo a versão do modelo, os dados de entrada e o prompt utilizado. Esses registros são armazenados no Amazon S3 e auditados via CloudWatch e CloudTrail.

Segurança e Governança

A segurança é tratada de forma robusta com autenticação centralizada via Auth0. Cada sessão gera um JWT que é validado no backend antes de qualquer operação. A AWS aplica políticas de IAM específicas para limitar o acesso aos dados de acordo com o perfil do usuário.

Na camada de visualização, o Amazon QuickSight aplica controle de acesso por linha (Row-Level Security), garantindo que os dashboards exibam apenas informações pertinentes ao usuário autenticado. Toda atividade é registrada, permitindo rastreamento completo de cada insight gerado e exibido.

Exemplo de Insight e Aplicação Prática

Um exemplo representativo do tipo de insight gerado seria:

> "Com base nos seus gastos médios dos últimos 6 meses, você pode economizar até R$ 250/mês ao consolidar seus investimentos em fundos isentos. Em setembro, sua previsão de gastos é de R$ 7.430, com destaque em viagens e alimentação."

Este texto foi produzido a partir das transações armazenadas no Redshift, cruzadas com documentos de fundos disponíveis na Oracle Cloud e regras de recomendação internas. O insight é apresentado em uma seção dedicada do aplicativo chamada "Insights Inteligentes", onde também são embutidos gráficos do Amazon QuickSight com a projeção de gastos futuros.

Avaliação Comparativa: Amazon Q Business vs AWS Bedrock

Durante a definição da arquitetura, foi considerada a possibilidade de utilizar tanto o Amazon Q Business quanto o AWS Bedrock como base para a IA generativa. Ambos os serviços oferecem vantagens relevantes, mas com características distintas.

O Amazon Q Business se destaca pela entrega rápida e baixo esforço de configuração, oferecendo integração nativa com IAM, RAG simplificado por meio de bases de conhecimento (Knowledge Bases) e conectores empresariais prontos. Ele proporciona rastreabilidade automática dos insights e pode ser incorporado facilmente em ambientes de produção com foco em governança e conformidade.

Por outro lado, o AWS Bedrock é mais indicado em cenários onde a personalização é um fator crítico. Ele permite o uso de múltiplos modelos de linguagem (Claude, Mistral, Llama, Titan etc.), com liberdade total para orquestrar o pipeline de RAG, criar prompts dinâmicos, e construir experiências customizadas de geração de conteúdo. Em contrapartida, sua adoção exige maior esforço de desenvolvimento e governança personalizada.

No contexto do NexoBank, optamos por iniciar a implementação com o Amazon Q Business, considerando sua rápida integração, baixo custo operacional e funcionalidades de rastreabilidade prontas para produção. Esta escolha não impede, no entanto, uma evolução futura para o AWS Bedrock caso o projeto demande maior flexibilidade na geração de texto, múltiplos modelos ou prompts mais complexos. Essa abordagem incremental garante velocidade de entrega no curto prazo, sem comprometer a escalabilidade e sofisticação futuras da solução.

Considerações Finais

A solução proposta combina o poder da IA generativa com práticas sólidas de arquitetura em nuvem, garantindo segurança, escalabilidade e rastreabilidade. A integração entre AWS e OCI foi cuidadosamente planejada para respeitar o legado da empresa, sem comprometer a experiência do cliente. A escolha pelo Amazon Q Business e QuickSight permite acelerar o time-to-market mantendo conformidade e controle sobre os dados.

Este projeto está alinhado com a missão do NexoBank de oferecer experiências personalizadas e inteligentes para seus clientes, utilizando o que há de mais moderno em tecnologias de nuvem e inteligência artificial.

Referências

AWS. (2025). *autenticação Okta na AWS*. Fonte: https://aws.amazon.com/pt/solutions/guidance/okta-phone-based-multi-factor-authentication-on-aws/

Oracle. (s.d.). *conectividade multicloud entre a Oracle Cloud Infrastructure (OCI) e a Amazon Web Services (AWS)* . Fonte: https://docs.oracle.com/pt-br/solutions/learn-about-multicloud-arch-framework/cloud-network-access1.html#GUID-99AD66BE-CD61-4D05-8CD3-24F7D66F41F8