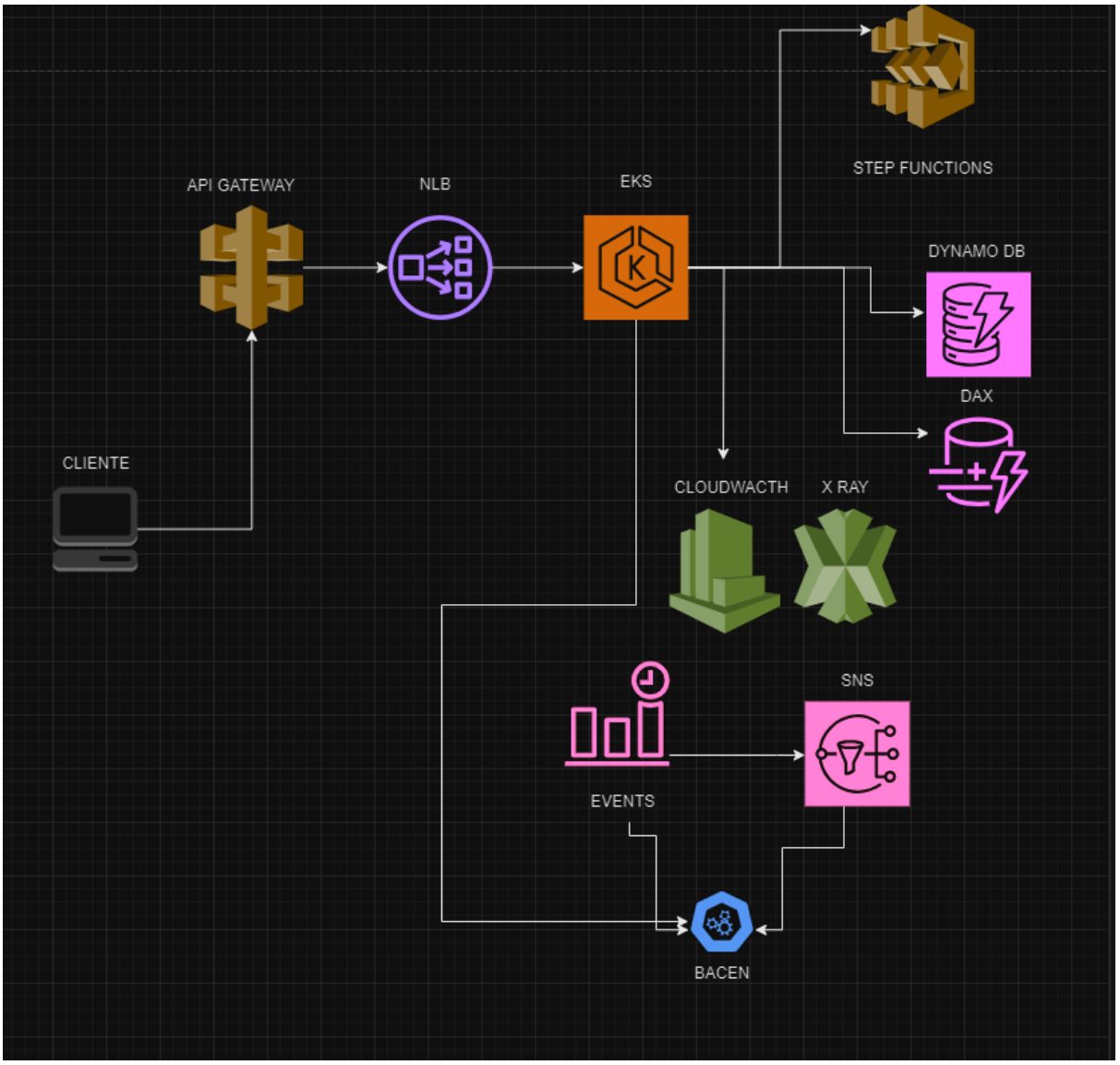
Desenho de Solução



Proposta de Escalonamento para Oscilação de Carga:
O Amazon Elastic Kubernetes Service (EKS) pode ser usado para escalar automaticamente com base na utilização da CPU ou em outras métricas personalizadas, garantindo que a aplicação possa lidar com picos de carga.

Proposta de Observabilidade:
Amazon CloudWatch para monitorar métricas de desempenho da aplicação, logs e eventos. Também o AWS X-Ray para rastrear e analisar o desempenho das chamadas de API, identificando gargalos e áreas de melhoria.

Escolha da Solução de Banco de Dados:
Amazon DynamoDB devido à sua escalabilidade automática, desempenho consistente e baixa latência de leitura/gravação. DynamoDB é capaz de lidar com o throughput necessário e oferece opções de replicação multi-região para alta disponibilidade

Justificativa para o Uso de Caching:
Amazon DynamoDB Accelerator (DAX):
Fornece baixa latência e alto desempenho para aplicativos que usam o DynamoDB. Oferece integração direta com o DynamoDB e atualizações de cache automáticas sempre que os dados no DynamoDB são modificados.

Suporte a Alto Throughput:
O Amazon API Gateway junto com o Network Load Balancer para distribuir o tráfego entre várias instâncias da API, usando o escalonamento do EKS horizontalmente

Estratégia em Caso de Falha de Dependências:
AWS Step Functions para coordenar a lógica de fallback e a resposta ao cliente, garantindo que o serviço permaneça disponível mesmo durante falhas de dependências. Para acionar a execução dos fluxos de trabalho do Step Functions quando ocorrerem eventos que exijam fallback, pode ser feito por meio de chamadas de API assíncronas, eventos de mensageria ou interceptores de exceção.

Estratégia em Caso de Throttling do BACEN:
O Amazon CloudWatch Events para monitorar o status de resposta da API em relação ao BACEN. Se o status 429 for detectado. O Amazon Simple Notification Service (SNS) pode ser acionado para notificar o órgão regulador, e ter uma estratégia de backoff para tentativas subsequentes de chamadas, garantindo uma abordagem para reenvio de solicitações.