**TESTE TÉCNICO DATA ENGINEER**

1. **Ingestão de Dados:** O primeiro passo é lidar com a ingestão de dados. Os dados de entrada são recebidos em formato JSON a partir de um serviço de mensageria. Para isso, podemos usar o Amazon Kinesis Data Streams ou o Amazon SQS, que é um serviço de streaming de dados em tempo real que pode capturar, processar e armazenar terabytes de dados por hora de várias fontes.
   1. **Descrição KDS**
      1. **Streaming em tempo real:** O KDS é projetado para fornecer baixa latência, proporcionando a capacidade de visualizar e analisar dados quase em tempo real. Isso é particularmente útil para aplicações que requerem respostas rápidas, como monitoramento de desempenho, análise de log em tempo real, entre outros.
      2. **Escalabilidade:** KDS é altamente escalável, permitindo que você inicie com um pequeno fluxo de dados e aumente conforme suas necessidades crescem. Você pode escalar facilmente a quantidade de shards (unidades básicas de capacidade do stream) em seu stream para acomodar mais produtores de dados ou consumidores. Cada shard pode lidar com até 1 MB de dados por segundo para gravação e até 2 MB por segundo para leitura.
      3. **Resiliência e Durabilidade:** Os dados no KDS são armazenados de forma redundante em vários servidores em várias zonas de disponibilidade dentro de uma região AWS para proporcionar alta disponibilidade e durabilidade. Os dados armazenados no KDS são retidos por padrão por 24 horas, mas podem ser configurados para serem retidos por até 7 dias.
      4. **Integração com outras ferramentas AWS:** KDS se integra bem com outras ferramentas da AWS, como AWS Lambda, Amazon Kinesis Data Analytics, Amazon Kinesis Data Firehose, Amazon S3, Amazon Redshift, Amazon EMR, AWS Glue e Amazon SageMaker, proporcionando um ecossistema completo para análise de dados em tempo real.
   2. **Integrações KDS**
      1. **Amazon S3:** O Kinesis Data Firehose pode ser usado para entregar dados do Kinesis Data Streams diretamente ao Amazon S3 para armazenamento durável e análise posterior.
      2. **AWS Lambda:** AWS Lambda pode ser usado para processar os dados do Kinesis Data Streams, permitindo transformações, filtragem, ou até mesmo a invocação de outros serviços AWS com base nos dados do stream.
      3. **Amazon SageMaker:** O Kinesis Data Streams pode ser integrado ao Amazon SageMaker para servir como entrada de dados para treinamento de modelos de machine learning.
      4. **AWS Glue:** AWS Glue pode ser usado para catalogar os dados do Kinesis Data Streams e torná-los disponíveis para ferramentas de análise como Athena e Redshift Spectrum.
2. **Processamento e Transformação de Dados:** O AWS Lambda ou AWS Glue ETL Spark será usado para processar os dados ingeridos. Nós podemos escrever uma função Lambda em Python que é acionada sempre que novos dados são enviados para o Kinesis Data Stream. Esta função Lambda transforma o formato JSON para o formato Parquet. O AWS Glue pode ser usado para definir os metadados dos dados e transformar o formato do arquivo.
   1. **Descrição Lambda**
      1. **Sem servidor:** O AWS Lambda é um serviço sem servidor. Isso significa que os desenvolvedores não precisam se preocupar com a infraestrutura de servidor na qual o código é executado. A AWS gerencia a infraestrutura para os usuários.
      2. **Orientado a eventos:** O AWS Lambda é projetado para ser acionado por eventos. Esses eventos podem vir de outras partes do ecossistema da AWS, como S3 (por exemplo, quando um arquivo é carregado), DynamoDB (quando um item é alterado), ou mesmo em resposta a alterações de dados em um banco de dados ou atualizações de estado de um dispositivo IoT.
      3. **Escala automática:** O AWS Lambda escala automaticamente. Seu aplicativo começa e escala automaticamente com o tamanho da sua carga de trabalho. Quando a carga de trabalho diminui, o Lambda reduz automaticamente a infraestrutura para economizar recursos.
      4. **Modelo de cobrança:** Com o AWS Lambda, você paga apenas pelo tempo de computação que você consome. Não há cobrança quando seu código não está sendo executado.
      5. **Linguagens de programação suportadas:** O AWS Lambda suporta várias linguagens de programação, incluindo Node.js, Java, C#, Go, PowerShell, Python e Ruby. Você também pode usar pacotes de tempo de execução personalizados para usar outras linguagens de programação.
   2. **Integrações KDS**
      1. **Amazon S3:** O Lambda pode ser usado para processar arquivos assim que são carregados no S3. Por exemplo, você pode configurar uma função Lambda para redimensionar imagens assim que forem carregadas em um determinado bucket S3.
      2. **Amazon DynamoDB:** O Lambda pode processar streams do DynamoDB para reagir a alterações em tempo real em uma tabela do DynamoDB. Por exemplo, quando um item é adicionado a uma tabela, uma função Lambda pode ser acionada para enviar uma notificação.
      3. **Amazon Kinesis e Amazon SQS:** Da mesma forma que o DynamoDB, o Lambda pode processar dados de streams do Kinesis ou de filas do SQS. Isso permite que o Lambda processe grandes volumes de dados em tempo real.

**Treinamento de Modelos de Machine Learning:** Aqui, podemos usar o Amazon SageMaker, que é um serviço totalmente gerenciado que fornece a capacidade de construir, treinar e implantar modelos de machine learning rapidamente. Podemos criar um script em Python no Amazon SageMaker usando o framework PySpark para treinar o modelo de Machine Learning. O script lê os dados .parquet do bucket do S3, treina o modelo e salva o modelo treinado de volta no S3. Podemos configurar o Amazon CloudWatch para iniciar este script a cada 4 semanas.

**Aplicação do Modelo em Tempo Real:** O modelo de Machine Learning treinado é aplicado a novos dados em tempo real usando o script **apply\_ml\_model.py**. Este script lê os dados do Amazon Kinesis Data Stream, aplica o modelo de Machine Learning e escreve o resultado de volta em um novo tópico do Kinesis Data Stream. Aqui, podemos usar o Amazon Kinesis Data Analytics para executar o script **apply\_ml\_model.py** em um aplicativo de streaming que lê dados do Kinesis Data Stream, aplica o modelo e grava os resultados em outro Kinesis Data Stream.

**Armazenamento de Dados:** Todos os dados (dados de entrada, dados transformados e resultados do modelo) podem ser armazenados em buckets do Amazon S3 para fins de backup e análise posterior. Além disso, o Amazon S3 é um armazenamento resiliente, como exigido pelos requisitos.

**Consumo de Dados:** O novo tópico do Kinesis Data Stream contendo os resultados do modelo pode ser consumido por várias APIs.